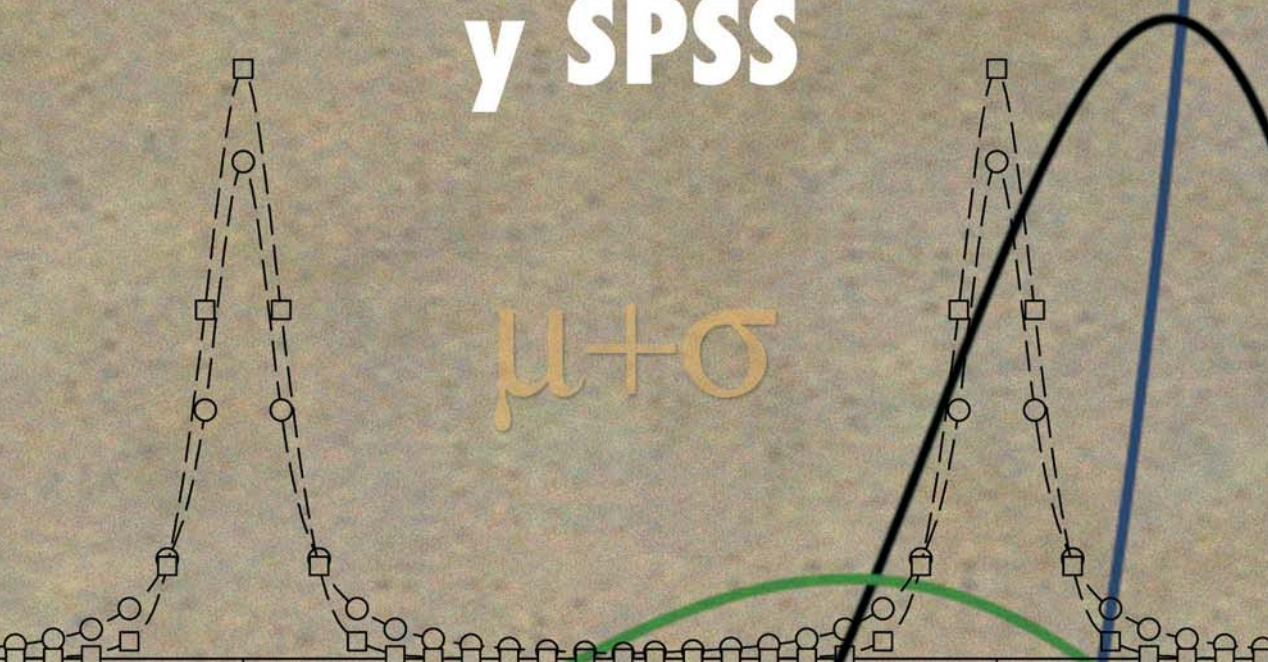
Cástor Guisande González
Antonio Vaamonde Liste
Aldo Barreiro Felpeto

Tratamiento de datos con R, STATISTICA y SPSS



DIAZ DE SANTOS

CÁSTOR GUISANDE GONZÁLEZ
Catedrático del área de Ecología. Universidad de Vigo (España)

ANTONIO VAAMONDE LISTE
Catedrático del área de Estadística e Investigación Operativa.
Universidad de Vigo (España)

ALDO BARREIRO FELPETO
Investigador del área de Ecología. Universidad de Vigo (España)

Tratamiento de datos con R, STATISTICA y SPSS

DIAZ DE SANTOS

©Cástor Guisande González, Antonio Vaamonde Liste y Aldo Barreiro Felpeto, 2011
(Versión papel)
© Cástor Guisande González, Antonio Vaamonde Liste y Aldo Barreiro Felpeto , 2013
(Versión electrónica)

Reservados todos los derechos.

Queda prohibida, salvo excepción prevista en la ley ,cualquier forma de reproducción, distribución, comunicación pública y transformación de esta obra sin contar con la autorización de los titulares de propiedad intelectual. La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (art.270 y siguientes del Código Penal). El Centro Español de Derechos Reprográficos (CEDRO) vela por el respeto de los citados derechos.

EdicionesDíazdeSantos
Email:ediciones@diazdesantos.es
www.editdiazdesantos.com

ISBN: 978-84-9969-612-6 (Libro electrónico)
ISBN:978-84-7978-998-5 (Libro en papel)

ÍNDICE

Prólogo y agradecimientos…… XV

I. REPRESENTACIÓN DE DATOS…… 1
I.1. COORDENADAS POLARES……..…… 1
I.1.1. Estandarización de los datos……..……....… 2
I.1.2. Asignación de ángulos a las variables…….…... 2
I.1.3. Representación de las coordenadas polares…… 7

II. ESTADÍSTICA DESCRIPTIVA…..……..…… 21
II.1. MEDIDAS DE POSICIÓN…….…….…… 21
II.1.1. Medidas de posición central……..… 21
II.1.1.1. Media aritmética……...… 21
II.1.1.2. Media geométrica…….. 22
II.1.1.3. Media armónica……...……..… 22
II.1.1.4. Moda……..…… 23
II.1.1.5. Mediana……..……... 23
II.1.1.6. Media acotada o recortada……..……. 25
II.1.1.7. Media ponderada…….. 25
II.1.2. Otras medidas de posición…… 27
II.2. MEDIDAS DE DISPERSIÓN……..…… 29
II.2.1. Amplitud…….…….. 30
II.2.2. Varianza y cuasivarianza…….……. 30
II.2.3. Desviación típica y cuasidesviación típica……... 30
II.2.4. Desviación absoluta respecto a la media y la mediana…….…….. 31
II.2.5. Coeficiente de variación……..…….. 33
II.2.6. Error estándar de la media…...…….. 33
II.2.7. Recorrido intercuartílico…….…….. 33
II.3. ESTADÍSTICA DESCRIPTIVA CON R……..... 33
II.4. ESTADÍSTICA DESCRIPTIVA CON SPSS……..… 37
II.5. ESTADÍSTICA DESCRIPTIVA CON STATISTICA 39

III. DISTRIBUCIÓN…… 47
III.1. TIPOS DE VARIABLES……..…….. 47
III.2. DISTRIBUCIONES PARA VARIABLES CONTINUAS…….……..…. 48

III.2.1. Normal……..… 48
III.2.1.1. Aplicaciones de la distribución Normal…. 49
III.2.1.2. Asimetría……..… 56
III.2.1.3. Apuntamiento o curtosis……..……. 57
III.2.2. *t* de Student……..…… 58
III.2.3. Chi-cuadrado……..…….. 60
III.2.4. *F* de Fisher-Snedecor……..…….. 61
III.3. DISTRIBUCIONES PARA VARIABLES DISCRETAS…….……..… 62
III.3.1. Binomial…….……. 62
III.3.2. Hipergeométrica…….…….…… 64
III.3.3. Poisson…….…… 66

IV. INTERVALOS DE CONFIANZA…….… 71
IV.1. INTERVALO DE CONFIANZA DE LA MEDIA DE UNA POBLACIÓN NORMAL…….……..….. 71
IV.1.1. Desviación típica conocida…….…. 71
IV.1.2. Desviación típica desconocida……..... 71
IV.1.2.1. Tamaño de muestra pequeño (< 30)…..…. 72
IV.1.2.2. Tamaño de muestra grande (≥ 30)…..…….. 72
IV.2. INTERVALO DE CONFIANZA DE LA DIFERENCIA DE MEDIAS DE DOS POBLACIONES NORMALES……...……..... 83
IV.2.1. Varianzas conocidas……..... 83
IV.2.2. Varianzas desconocidas……..…… 83
IV.2.2.1. Tamaño de muestra grande (≥ 30)……..… 83
IV.2.2.2. Tamaño de muestra pequeño (< 30)…...… 83
IV.2.2.2.1 Varianzas iguales…….…. 84
IV.2.2.2.2. Varianzas diferentes…….……..….. 84
IV.3. INTERVALO DE CONFIANZA DE LA VARIANZA Y DESVIACIÓN TÍPICA DE UNA POBLACIÓN NORMAL……. 91
IV.4. INTERVALO DE CONFIANZA DE LA RAZÓN DE VARIANZAS DE DOS POBLACIONES NORMALES…… 94

V. CONTRASTE DE HIPÓTESIS…….……... 103
V.1. TIPOS DE HIPÓTESIS…….……... 103
V.2. ESTADÍSTICO DE CONTRASTE……..…… 103
V.3. TIPOS DE ERRORES……..…… 103
V.4. REGIONES CRÍTICAS Y DE ACEPTACIÓN…...… 104

V.5. TIPOS DE CONTRASTES…….. 106

VI. CONTRASTES DE BONDAD DE AJUSTE……..… 107
VI.1. CHI-CUADRADO……..… 107
VI.2. TEST DE KOLMOGOROV-SMIRNOV……... 108
VI.3. TEST DE SHAPIRO-WILK …….. 108

VII. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUALITATIVAS……... 135
VII.1. MUESTRAS INDEPENDIENTES…….……..…… 135
VII.1.1. Variables politómicas…….…… 135
VII.1.1.1. Chi-cuadrado……..… 136
VII.1.1.2. Razón de verosimilitud (test G)……..… 136
VII.1.2. Variables dicotómicas…… 150
VII.1.2.1. Corrección de Yates…… 150
VII.1.2.2. Prueba de Fisher……..…….. 150
VII.2. MUESTRAS RELACIONADAS……..… 161
VII.2.1. Variables politómicas…….… 161
VII.2.1.1. ANOVA de Friedman……. 161
VII.2.2. Variables dicotómicas……..... 167
VII.2.2.1. Prueba de McNemar……..… 167
VII.2.2.2. Q de Cochran…….…….… 179

VIII. CONTRASTES DE INDEPENDENCIA Y ASOCIACIÓN EN VARIABLES CUALITATIVAS……. 189
VIII.1. VARIABLES POLITÓMICAS……..…….. 189
VIII.1.1. Pruebas de independencia…….… 189
VIII.1.1.1. Chi-cuadrado de Pearson y test G de razón de verosimilitud..... 189
VIII.1.2. Medidas de asociación……..…… 189
VIII.1.2.1. V de Cramer……..… 189
VIII.1.2.2. Coeficiente de Contingencia…….… 190
VIII.1.2.3. Coeficiente de Incertidumbre…… 190
VIII.2. VARIABLES DICOTÓMICAS……...… 202
VIII.2.1. Pruebas de independencia…….… 202
VIII.2.1.1. Corrección de Yates y prueba de Fisher... 202
VIII.2.1.2. Pruebas de Cochran y Mantel-Haenszel... 202
VIII.2.2. Medidas de asociación…….. 203
VIII.2.2.1. Phi (ϕ)……..…… 203
VIII.2.2.2. Riesgo relativo, razón de ventajas (odds ratio) y las pruebas de Breslow-Day y de Tarone.. 203

IX. CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS…… 217
IX.1. PRUEBAS PARAMÉTRICAS……...…… 217
IX.1.1. Requisitos…..……...…… 217
IX.1.2. Transformaciones…… 218
IX.1.3. t-test…….…… 218
IX.1.3.1. Muestras independientes……..…… 218
IX.1.3.2. Muestras dependientes……..… 240
IX.1.4. Análisis de varianza……. 251
IX.1.5. Análisis de covarianza…….…… 289
IX.1.6. Análisis de varianza y covarianza con medidas repetidas…….…….. 318
IX.1.7. Análisis de varianza anidado……...…… 363
IX.2. PRUEBAS NO PARAMÉTRICAS…… 386
IX.2.1. Contrastes para dos muestras independientes..… 386
IX.2.1.1. Prueba U de Mann-Whitney…… 386
IX.2.1.2 Test de rachas de Wald-Wolfowitz……...… 387
IX.2.1.3. Prueba de Kolmogorov-Smirnov para dos muestras…….. 388
IX.2.2. Contrastes para *k*-muestras independientes……. 400
IX.2.2.1. Contraste de la mediana……...……...… 400
IX.2.2.2 ANOVA de Kruskal-Wallis…...…….…… 400
IX.2.3. Contrastes para dos muestras dependientes…..… 408
IX.2.3.1. Contraste de los signos…… 408
IX.2.3.2. Prueba de Wilcoxon para pares relacionados……... 408
IX.2.4. Contrastes para *k*-muestras dependientes…… 415

X. CONTRASTES DE INDEPENDENCIA Y ASOCIACIÓN EN VARIABLES CUANTITATIVAS……... 421
X.1. CORRELACIONES BIVARIADAS SIMPLES…...… 421
X.1.1. Coeficiente de correlación de Pearson (r)…… 421
X.1.2. Coeficiente de correlación de Spearman (ρ)…… 422
X.1.3. Coeficiente τ de Kendall…….……..… 423
X.1.4. Coeficiente Gamma (γ)……...…….. 424
X.2. CORRELACIONES BIVARIADAS PARCIALES…. 439
X.2.1. Coeficiente de correlación parcial de Pearson..… 439
X.2.2. Coeficiente de correlación parcial de Kendall ($T_{xy.z}$)……..…….. 440
X.3. CORRELACIÓN MÚLTIPLE…….…… 449
X.3.1.Coeficiente de correlación múltiple de Pearson.… 449

X.3.2. Coeficiente de concordancia de Kendall (*W*)…… 450

XI. REGRESIONES ……..…… 453
XI.1 MODELOS DE REGRESIÓN PARA VARIABLES DEPENDIENTES CUANTITATIVAS…….. 453
XI.1.1.Requisitos……. 453
XI.1.2. Regresión simple…….…… 455
XI.1.3. Regresión múltiple……...…… 486
XI.1.4. Otras regresiones simples o múltiples no lineales……. 519
XI.1.4.1. Curva logística……. 519
XI.1.4.2. Curva de crecimiento de von Bertalanffy… 520
XI.1.4.3. Curva de crecimiento de Gompertz……. 521
XI.1.4.4. Relación entre tasas y variables…..……... 521
XI.2 MODELOS DE REGRESIÓN PARA VARIABLES DEPENDIENTES CUALITATIVAS…….……. 537
XI.2.1. Regresión logística binomial……... 537
XI.2.2. Regresión logística multinomial…….. 557

XII. SERIES TEMPORALES…… 585
XII.1. MODELO ARIMA…… 585
XII.1.1. Autocorrelación……..…… 585
XII.1.2. Series estacionarias o no estacionarias…...…… 586
XII.1.3. Estacionalidad…… 586
XII.1.4. Covariables…….… 586
XII.1.5. Construcción del modelo…… 586
XII.2. CICLOS……..…… 637

XIII. ANÁLISIS MULTIVARIANTE: MÉTODOS FACTORIALES…….. 667
XIII.1. ANÁLISIS DE COMPONENTES PRINCIPALES. 667
XIII.2. ANÁLISIS FACTORIAL…….… 697
XIII.3. ANÁLISIS DE CORRESPONDENCIAS……...…. 719
XIII.4. ESCALAMIENTO MULTIDIMENSIONAL…..... 743

XIV. ANÁLISIS MULTIVARIANTE: MÉTODOS DE CLASIFICACIÓN…….……. 759
XIV.1. CLASIFICACIÓN JERÁRQUICA……..… 759
XIV.1.1. Criterio de distancia……..…… 760
XIV.1.2. Criterio de aglomeración…….. 760
XIV.1.3. Dendrograma…..…….……. 761

XIV.2. CLASIFICACIÓN DE *K*-MEDIAS……..…… 773
XIV.3. ANÁLISIS DISCRIMINANTE……..……. 787
XIV.3.1. Hipótesis……..……. 787
XIV.3.2. Método paso a paso……..…… 787
XIV.3.3. Influencia de las distintas variables independientes…….. 788
XIV.3.4. Validación…….…… 788

XV. MODELOS DE SIMULACIÓN…… 811
XV.1. EL USO DE MODELOS…….…….. 811
XV.2. PASOS A CONSIDERAR PARA EL DESARROLLO DE UN MODELO…….……..... 812
XV.3. INTRODUCCIÓN AL MODELADO CON EL PROGRAMA STELLA……..……... 813
XV.3.1. Significado de los iconos y menús específicos de la barra de herramientas en niveles *Model*, *Map* e *Interface*..... 814
XV.3.2. Ejemplo de funcionamiento de los iconos y menús básicos ……....…… 819
XV.4. EJEMPLO DE MODELADO CON EL PROGRAMA STELLA: DESARROLLO CONCEPTUAL Y MANEJO PRÁCTICO…….…… 832
XV.4.1. Dinámica poblacional de la presa……..… 833
XV.4.2. Dinámica poblacional del depredador y su influencia sobre la presa…….… 834
XV.4.3. Influencia de la explotación sobre la población del depredador…….……...……..… 836

BIBLIOGRAFÍA……...……. 839

GUÍA RESUMEN……..…… 845

APÉNDICE I. TABLAS ESTADÍSTICAS…..……..…… 851
Tabla 1. Áreas bajo la curva Normal estándar……..… 851
Tabla 2. Valores críticos de la distribución *t* de Student..… 852
Tabla 3. Valores críticos de la distribución χ^2……..… 853
Tabla 4. Valores críticos de la distribución *F* de Fisher-Snedecor……... 861
Tabla 5. Valores críticos del estadístico de Kolmogorov-Smirnov…… 873
Tabla 6. Valores críticos del estadístico de Lilliefors…….. 874

Tabla 7. Distribución del estadístico de contraste de Durbin-Watson…… 875

APÉNDICE II. INSTALACIÓN Y CONCEPTOS BÁSICOS SOBRE R ……..…… 877
AII.1. SOBRE R…….…….. 877
AII.1.1. ¿Qué es R?..... 877
AII.1.2. Instalación……..…… 877
AII.1.3. Interfaz……... 878
AII.2. PRIMEROS PASOS…..……..… 879
AII.2.1. Entrar/salir de R y seleccionar directorio de trabajo…….……. 879
AII.2.2. Crear un archivo script……...…… 880
AII.2.3. Instalar y cargar paquetes y conjuntos de datos. 882
AII.2.4. Actualización de R…….…… 884
AII.2.5. Ayuda, ejemplos y demos……..…… 887
AII.2.6. Documentación sobre R …… 890
AII.3. R COMMANDER …… 890
AII.3.1. Características de R Commander……...…… 890
AII.3.2. El menú de R Commander…….…… 893
AII.3.2.1. Fichero…… 893
AII.3.2.2. Editar……..…… 893
AII.3.2.3. Datos……...…… 894
AII.3.2.4. Estadísticos…….…… 895
AII.3.2.5. Gráficas…..……..…… 895
AII.3.2.6. Modelos……..…… 896
AII.3.2.7. Distribuciones…..…… 896
AII.3.2.8. Herramientas…..……...…..…… 897
AII.3.2.9. Ayuda…….…… 897
AII.4. EL LENGUAJE R…….……. 897
AII.4.1. Apuntes preliminares…….…… 897
AII.4.2. Importación de datos……..…… 900
AII.4.3. Guardar datos y generar archivos de salida…… 903
AII.4.4. Objetos……... 906
AII.4.5. Operadores……...…….……. 909
AII.4.6. Funciones……... 909
AII.4.6.1. Funciones built in……..…. 909
AII.4.6.2. Funciones definidas por el usuario…… 914
AII.4.7. Bucles, ciclos o *loops*……...…….. 918
AII.4.8. Vectores…….……. 920
AII.4.8.1. Creación de vectores…….…….. 920

AII.4.8.2. Trabajando con vectores……. 922
AII.4.9. Matrices y *Arrays*…….....……...…….. 924
AII.4.9.1. Creación de matrices y Arrays…….…….. 924
AII.4.9.2. Trabajando con matrices y Arrays…..…… 929
AII.4.10. Listas…..……...…… 933
AII.4.10.1. Creación de listas…….…… 933
AII.4.10.2. Trabajando con listas……...…… 934
AII.4.11. *Data frames* ……..…… 936
AII.4.11.1. Creación de data frames……..……..…… 936
AII.4.11.2. Seleccionando variables y casos en data frames……. 939
AII.4.11.3. Funciones importantes en data frames…. 944
AII.4.12. Editor de datos……..…… 948
AII.4.13. Texto ……..…….. 948
AII.4.14. Gráficos ……..…….. 950
AII.4.14.1. Aspectos generales..…….…… 950
AII.4.14.2. Gráfico de dispersión……..……..…. 952
AII.4.14.3. Representar modelos y gráficos en paneles…… 957
AII.4.14.4. Diagrama de cajas…….……... 958
AII.4.14.5. Graficar curvas de ajuste y caracteres especiales……...…… 960
AII.4.14.6. Histograma…….…….. 961
AII.4.14.7. Gráficos de barras…….……... 965
AII.4.14.8. Gráficos de contornos y superficies…..… 969
AII.4.14.9. Configuración de márgenes y ejes adicionales……. 971
AII.4.14.10. Formas especiales……... 972

ÍNDICE DE CONCEPTOS…… 975

PRÓLOGO Y AGRADECIMIENTOS

Los investigadores necesitan la Estadística para obtener conclusiones válidas a partir de los datos. El método estadístico se ha convertido en parte esencial de la generación de conocimiento científico, y en prácticamente todas las publicaciones especializadas –en casi cualquier campo de conocimiento– las técnicas estadísticas tienen un papel muy importante, hasta el punto de que las revistas científicas suelen disponer de revisores con amplios conocimientos en este campo.

Aunque los usuarios de los métodos de tratamiento y análisis de datos tienen en general conocimientos de Estadística, no suelen ser expertos; para resolver problemas específicos tienen dos opciones: consultar con un especialista –siempre recomendable si el problema es complejo– o recurrir a algún manual de consulta. Frecuentemente los libros de Estadística son excesivamente teóricos, prestando atención únicamente a los aspectos matemáticos, demostraciones, teoremas y propiedades (sin duda de gran importancia para conocer en profundidad las técnicas estadísticas), o bien libros que omiten totalmente los principios básicos y el esquema conceptual que permite comprender el funcionamiento de cada técnica estadística, para centrar todos sus esfuerzos en la aplicación en si misma, presentada como receta al alcance del lector profano.

Los libros más teóricos, pese a su corrección formal, resultan de escasa utilidad para el usuario normal de las técnicas estadísticas, ya que se ve obligado a convertirse en un experto, o al menos a intentarlo, dedicando a ello más tiempo del que dispone, para poder llegar a resolver sus problemas inmediatos de tratamiento de datos, a menudo sin lograrlo. Los libros más aplicados, por el contrario, parecen muy adecuados para el lector que tiene un problema estadístico similar al descrito en el capítulo correspondiente, pero pronto averigua que lamentablemente los problemas son siempre distintos, que existen cuestiones que la receta no trata o no resuelve, y que al utilizarla nunca está a salvo de cometer errores insalvables, que invalidarán los resultados de su investigación.

El principal objetivo de este libro es servir de ayuda a los investigadores, cualquiera que sea su ámbito de conocimiento, que necesitan utilizar técnicas estadísticas para tratar y analizar sus datos. Con un enfoque centrado en la aplicación, se hace énfasis en la explicación de cada método estadístico con un lenguaje accesible, tratando de aclarar de forma breve pero razonada los principios básicos de cada técnica. Numerosas aplicaciones con casos reales son utilizadas para fijar los conceptos y explicar los pasos a seguir en el tratamiento de los datos, desde la introducción de datos hasta la interpretación de los resultados obtenidos.

Se utilizan tres programas estadísticos: SPSS, STATISTICA, y R. Todos ellos son programas de ámbito general muy conocidos y utilizados para el tratamiento estadístico de datos en diversos campos. R es un programa estadístico de uso libre y gratuito, que recientemente ha alcanzado una gran difusión. A la ventaja –nada desdeñable– de su utilización gratuita, une una potencia de análisis superior a la de cualquier otro programa de carácter general, junto con un ritmo de desarrollo de aplicaciones que permite adivinar que pronto desplazará del mercado a los programas más costosos. R es un proyecto abierto y participativo en el

que colaboran –de forma libre y espontánea– miles de investigadores estadísticos, lo que asegura el debate, la crítica, y la actualización permanente de los procedimientos; su único inconveniente, por el momento, es el entorno o interfase de aplicación, con menús de usuario poco (o nada) desarrollados en algunas técnicas, que utiliza un lenguaje de programación que puede parecer complejo a los no iniciados.

Todas las técnicas son aplicadas con los tres programas. Para ello se utilizan datos distintos, aunque a veces el mismo ejemplo es resuelto con dos programas diferentes para que el lector pueda comparar y deducir sus ventajas y limitaciones. Se trata de conseguir que el lector pueda reproducir con sus propios datos cualquier aplicación y, por ello, se sigue siempre un esquema de pasos sucesivos, con el que debe resultar razonablemente sencillo completar un nuevo análisis mediante cualquiera de los tres programas estadísticos mencionados.

Todos los ejemplos se explican detalladamente, para que puedan ser repetidos incluso por el usuario que no tiene conocimientos previos de Estadística ni del programa. Las distintas técnicas estadísticas son presentadas de forma independiente, de manera que el lector interesado en cualquiera de ellas pueda empezar directamente por el capítulo correspondiente.

Los supuestos o hipótesis que cada método requiere son verificados cuidadosamente: es necesario que el lector compruebe también todas las hipótesis en cualquier aplicación de la misma técnica, y lo haga constar así junto a la presentación de los resultados. Los métodos estadísticos solo son válidos cuando se aplican correctamente en su contexto adecuado, y es muy frecuente encontrar aplicaciones desafortunadas en las que obviamente no se cumplen las condiciones necesarias, o en las que simplemente no se han comprobado.

Hemos prestado también especial atención a la interpretación correcta de los resultados obtenidos. Es bastante habitual observar –en la publicación de resultados– cómo se obtienen conclusiones que no pueden deducirse de los datos, generalmente porque se interpreta de forma laxa o generalizada el resultado de una prueba estadística, porque se ignoran las características de los datos utilizados, o porque se desconocen las posibilidades y limitaciones del método estadístico. Aunque en este libro se intenta explicar con claridad cómo deben interpretarse correctamente los distintos resultados, lo cierto es que éste es el aspecto más difícil para el usuario no experto. El único antídoto frente al mal uso de la Estadística consiste en un conocimiento profundo de las técnicas, por lo que recomendamos vivamente que el usuario consulte a un especialista cuando tenga dudas de interpretación en la aplicación de métodos estadísticos complejos.

El libro se acompaña de un CD con los archivos de datos de todos los ejemplos utilizados, así como de los programas de R necesarios para su análisis, con el fin de facilitar su utilización. Una breve guía resumen al final del libro puede utilizarse para una primera orientación sobre la técnica estadística que debe ser utilizada, en función del tipo de datos del que se dispone y de los objetivos buscados con el análisis.

Esta obra no tiene la pretensión de ser exhaustiva. Muchas técnicas simples de Estadística Descriptiva y algunos métodos estadísticos de carácter más específico no son tratados para no aumentar excesivamente su extensión. Hemos tratado de incluir la mayoría de las técnicas de tratamiento estadístico de datos que los investigadores necesitan en su trabajo habitual, añadiendo la resolución de problemas frecuentes que no tienen en otros textos una consideración suficiente, y hemos procurado también situarnos en el lugar del usuario para dar una respuesta satisfactoria a sus demandas, de modo que pueda superar sin dificul-

tades las barreras que a menudo impiden la aplicación de esta herramienta imprescindible que es la Estadística.

Por último, queremos agradecer al profesor Kenneth Roy Cabrera Torres de la Universidad Nacional de Colombia Sede Medellín, su ayuda con los *scripts* de R en los primeros capítulos de este libro.

Capítulo I

REPRESENTACIÓN DE DATOS

I.1. Coordenadas polares

Un paso obligado previo a la realización de cualquier tratamiento de datos es representar los datos gráficamente. Esto es necesario por muchos motivos: por ejemplo, para ver el tipo de relación que existe entre dos variables (lineal, logarítmica, exponencial, etc.); para identificar posibles "*outliers*", datos que son muy diferentes del resto y que se pueden deber simplemente a que hemos introducido mal los datos en el ordenador; para ver el tipo de distribución y variabilidad de los datos, etc. Existen muchos tipos de gráficas que permiten hacer representaciones, que son bien conocidas y de uso común por la mayoría de las personas que trabajan con datos. En este capítulo solo vamos a tratar un tipo de representación que no es tan conocido, el gráfico de coordenadas polares.

Un problema frecuente que surge a la hora de mostrar gráficamente los resultados obtenidos es que es necesario representar más de dos ejes de coordenadas. Sin embargo, en un plano bidimensional lo máximo que se puede dibujar son tres ejes. Las coordenadas polares permiten representar en un gráfico bidimensional cualquier número de ejes de coordenadas.

Supongamos que se quiere representar el nicho de varias especies de árboles considerando como variables del nicho el pH medio del suelo en el que aparece la especie, la temperatura media, la salinidad media del suelo, la pluviosidad media y la concentración media de nitratos en el suelo. Las coordenadas polares consideran las diferentes variables como vectores de fuerza, de tal forma que las coordenadas polares *X* e *Y* de cada especie serían el punto de equilibrio de todos los vectores es decir, de todas las variables (Figura I.1).

El módulo de cada vector sería el valor de la variable y el ángulo de cada vector tendría un valor convenido asignado a cada variable. Por lo tanto, cada especie tendría unas coordenadas *X* e *Y*, lo cual permite su representación en un plano bidimensional y, además, estas coordenadas *X* e *Y* vendrían determinadas por el efecto conjunto de todas las variables consideradas para definir el nicho de las especies. Es decir, cada especie ocuparía una posición en el plano en función de las variables del nicho, lo cual permitiría ver gráficamente que especies tienen un nicho más parecido.

Debido a que la representación polar se basa en vectores de fuerza, solo se puede utilizar con variables que tengan igual peso, es decir, que los valores de las variables puedan ser comparables. En el caso de que las variables no sean equiparables es necesario realizar una estandarización previa de los datos.

En un análisis multivariante de componentes principales, las puntuaciones obtenidas en cada eje para cada caso serían un ejemplo de variables que son en general equiparables. Los porcentajes de alimento consumido por varias especies también serían otro caso de variables equiparables, así como, por ejemplo, el porcentaje granulométrico del suelo.

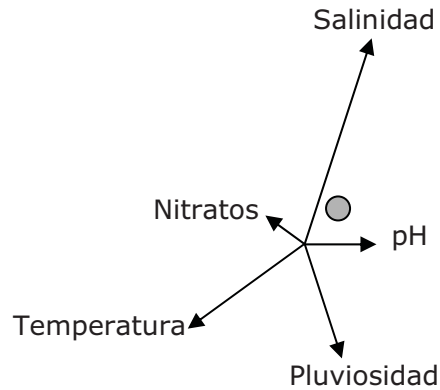


Figura I.1 Esquema de un sistema de coordenadas polares donde el círculo gris indica el centro de gravedad o punto de equilibrio de todas las variables. Las coordenadas de ese punto gris serían las coordenadas *X* e *Y* de la especie.

Por el contrario, si por ejemplo queremos una representación de varias lagunas, para ver gráficamente qué lagunas tienen unas características físico-químicas semejantes, considerando la temperatura, conductividad, pH, oxígeno, etc., estas variables no serían equiparables y sería necesario realizar una estandarización previa.

I.1.1. ESTANDARIZACIÓN DE LOS DATOS

Como ejemplo de estandarización previa al cálculo de las coordenadas polares vamos a trabajar con datos físico-químicos de distintas lagunas, para las que, en algunas de ellas, existen varias estaciones de muestreo.

En la Tabla I.1 (Cuadro I.1) se muestran los datos y en el archivo **Cuadro I.1.xls**, además de los datos, se muestran todos los pasos a seguir para estandarizar los datos, los cuales también se describen en el Cuadro I.1.

En el Cuadro I.2 hay otro ejemplo a modo de práctica, del método para estandarizar datos.

I.1.2. ASIGNACIÓN DE ÁNGULOS A LAS VARIABLES

Un paso previo al cálculo de las coordenadas es asignar ángulos a las variables consideradas. La asignación de ángulos a las variables es diferente si hay valores negativos (por ejemplo una salida de un análisis multivariante) o si, por el contrario, solo hay valores estandarizados en una escala de 0 a 1 y, por tanto, no hay valores negativos.

En el caso de que solo existan valores positivos en una escala de 0 a 1 (los valores están estandarizados), lo que se hace es restar a todos los valores 0,5 antes de comenzar con el proceso de asignar ángulos a las variables. Si el rango de los datos no es de 0 a 1 (los valores no están estandarizados) ya que, por ejemplo, es una salida de un análisis multivariante que tiene valores negativos, no se resta el valor de 0,5 a todos los datos.

CUADRO I.1. Estandarización de datos

EJEMPLO. La estandarización de los datos físico-químicos de diferentes lagunas se muestra en la siguiente tabla. Los datos están en el archivo **Cuadro I.1.xls**.

Tabla I.1. Datos físico-químicos de distintas lagunas.

Laguna	Estación	Temperatura (ºC)	Conductividad (μS cm^{-1})	pH	Nitrato (μM)	Nitrito (μM)	Amonio (μM)	Fosfato (μM)	Silicato (μM)
1	1	29,1	495	6,08	0,00	1,090	1,140	0,320	62,78
2	1	28,1	1537	7,85	0,52	0,280	0,940	0,384	231,67
	2	28,6	1568	8,52	0,00	0,253	0,839	0,428	262,31
3	1	28,2	755	8,52	0,52	0,275	0,937	0,384	231,67
	2	28,5	739	8,24	0,00	0,348	0,339	0,761	172,69
	3	28,6	748	8,60	0,13	0,386	0,419	0,867	189,41
4	1	27,5	905	7,63	0,00	0,893	0,915	1,056	91,699
5	1	26,4	419	7,72	0,00	0,708	0,668	0,258	113,18
6	1	23,9	1034	7,64	0,42	0,301	1,229	0,263	189,04
7	1	32,6	217	7,42	0,30	0,843	2,076	0,618	118,00
8	1	31,3	371	7,32	0,00	0,697	0,635	0,489	99,225
9	1	32,3	1162	7,17	0,04	0,248	0,762	1,811	326,63

Paso 1. Cálculo de los valores máximos y mínimos de las variables.

	Temperatura (ºC)	Conductividad (μS cm^{-1})	pH	Nitrato (μM)	Nitrito (μM)	Amonio (μM)	Fosfato (μM)	Silicato (μM)
Máximo	32,6	1568	8,6	0,52	1,09	2,076	1,811	326,63
Mínimo	23,9	217	6,08	0	0,248	0,339	0,258	62,78

Paso 2. Estandarización a una escala de 0 a 1 de todas las variables.

A cada uno de los valores de las variables se aplica la siguiente fórmula:

$$VE = \frac{x - Min}{Max - Min}$$

donde *VE* es el valor estandarizado, *Max* y *Min* son los valores máximo y mínimo de cada variable, respectivamente, que se calcularon en el paso 1, y *x* es cada uno de los valores de cada variable. Los valores estandarizados se muestran en la Tabla I.2.

Tabla I.2. Valores estandarizados de los datos físico-químicos que se muestran en la Tabla I.1.

Laguna	Estación	Temperatura (ºC)	Conductividad (μS cm^{-1})	pH	Nitrato (μM)	Nitrito (μM)	Amonio (μM)	Fosfato (μM)	Silicato (μM)
1	1	0,598	0,206	0,000	0,000	1,000	0,461	0,040	0,000
2	1	0,483	0,977	0,702	1,000	0,038	0,346	0,081	0,640
	2	0,540	1,000	0,968	0,000	0,006	0,288	0,109	0,756
3	1	0,494	0,398	0,968	1,000	0,032	0,344	0,081	0,640
	2	0,529	0,386	0,857	0,000	0,119	0,000	0,324	0,417
	3	0,540	0,393	1,000	0,250	0,164	0,046	0,392	0,480
4	1	0,414	0,509	0,615	0,000	0,766	0,332	0,514	0,110
5	1	0,287	0,150	0,651	0,000	0,546	0,189	0,000	0,191
6	1	0,000	0,605	0,619	0,808	0,063	0,512	0,003	0,479
7	1	1,000	0,000	0,532	0,577	0,707	1,000	0,232	0,209
8	1	0,851	0,114	0,492	0,000	0,533	0,170	0,149	0,138
9	1	0,966	0,699	0,433	0,077	0,000	0,244	1,000	1,000

CUADRO I.1. *(Continuación)*

Estadística descriptiva con R

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro I.1.R**, el cual es un archivo *script* que contiene las instrucciones para realizar la estadística descriptiva que se muestra en este ejemplo. Este archivo está en el CD que acompaña al libro en la carpeta Capítulo I. Para grabar los resultados, es mejor pasar el archivo **Cuadro I.1.R** a un directorio del disco duro y seleccionarlo como directorio de trabajo.

Paso 1. Para importar bases de datos desde Excel a R se recomienda guardar el archivo de Excel en formato ".csv" y luego leer desde R utilizando la función read.csv o read.csv2 según el tipo de formato ".csv" usado por Excel para guardar los datos. Las instrucciones para realizar este tipo de importación de datos están en el Apéndice II en el apartado AII.4.2. En este ejemplo usaremos el archivo **Cuadro I.1.V.csv**, en el cual se encuentran los datos mostrados anteriormente en la Tabla I.1.

Paso 2. La siguiente ventana muestra las instrucciones del archivo *script* **Cuadro I.1.R**. Lo único que hay que hacer es cambiar la sección modificable por el usuario, para poder hacer la estandarización de cualquier tipo de archivo de datos.

```
# Cuadro I.1
# Estandarización de datos

#########################################################
# Sección modificable por el usuario
#########################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro I.1.V.csv",encoding="latin1")

# Selección de las variables de interés
varInteres<-c("Temperatura","Conductividad","pH","Nitrato",
              "Nitrito","Amonio","Fosfato","Silicato")

# Selección de las variables de identificación

varID<-c("Laguna","Estación")

# Nombre del archivo de salida con los resultados de la
# estandarización
nomSalida<-"Salida Cuadro I.1.V.csv"
```

Paso 3. En primer lugar hay que importar los datos y, por tanto, especificar el nombre del archivo de entrada, en este caso **Cuadro I.1.V.csv**.

datos<-read.csv2("Cuadro I.1.V.csv",encoding="latin1")

CUADRO I.1. *(Continuación)*

Paso 4. Una vez importados los datos hay que definir las variables que se quieren estandarizar.

varInteres<-c("Temperatura","Conductividad",”pH”,"Nitrato", "Nitrito","Amonio","Fosfato","Silicato")

Paso 5. A continuación se seleccionan las variables que se utilizan como códigos de identificación de los casos.

varID<-c("Laguna","Estación")

Paso 6. El último paso que tiene que hacer el usuario es definir el nombre del archivo donde se grabarán los valores estandarizados.

nomSalida<-"Salida Cuadro I.1.V.csv"

Paso 7. Para ejecutar el archivo *script es* posible posicionarse en cada línea del archivo *script* y pulsar *ctrl-R*, o seleccionar varias líneas o todo el archivo con *ctrl-A* y luego pulsar *ctrl-R*, para ejecutar varias líneas o todo el archivo a la vez. Los resultados se muestran en la siguiente ventana.

```
> resultado<-rbind(maximos,minimos)
>
> # Estandarización a una escala de 0 a 1 de todas las variables dada
> resultado2<-cbind(baseID,t((t(baseVar)-minimos)/(maximos-minimos)))
> if(length(varID)==1) colnames(resultado2)[1]<-varID
>
> ##########################################################
> # Sección que muestra los resultados
> ##########################################################
>
> # Muestra los resultados de mínimos y máximos
> resultado
        Temperatura Conductividad   pH Nitrato Nitrito Amonio Fosfato Silicato
maximos        32.6          1568 8.60    0.52    1.09   2.08    1.81   326.63
minimos        23.9           217 6.08    0.00    0.25   0.34    0.26    62.78
>
> # Muestra los resultados de la estandarización en la escala 0 a 1.
> resultado2
   Laguna Estación Temperatura Conductividad        pH    Nitrato    Nitrito
1       1        1   0.5977011     0.2057735 0.0000000 0.00000000 1.00000000
2       2        1   0.4827586     0.9770540 0.7023810 1.00000000 0.03571429
3       2        2   0.5402299     1.0000000 0.9682540 0.00000000 0.00000000
4       3        1   0.4942529     0.3982235 0.9682540 1.00000000 0.03571429
5       3        2   0.5287356     0.3863805 0.8571429 0.00000000 0.11904762
6       3        3   0.5402299     0.3930422 1.0000000 0.25000000 0.16666667
7       4        1   0.4137931     0.5092524 0.6150794 0.00000000 0.76190476
8       5        1   0.2873563     0.1495189 0.6507937 0.00000000 0.54761905
9       6        1   0.0000000     0.6047372 0.6190476 0.80769231 0.05952381
10      7        1   1.0000000     0.0000000 0.5317460 0.57692308 0.70238095
11      8        1   0.8505747     0.1139896 0.4920635 0.00000000 0.53571429
12      9        1   0.9655172     0.6994819 0.4325397 0.07692308 0.00000000
       Amonio    Fosfato  Silicato
1  0.45977011 0.03870968 0.0000000
2  0.34482759 0.07741935 0.6400985
3  0.28735632 0.10967742 0.7562251
4  0.34482759 0.07741935 0.6400985
5  0.00000000 0.32258065 0.4165624
6  0.04597701 0.39354839 0.4799318
7  0.33333333 0.51612903 0.1096077
8  0.18965517 0.00000000 0.1910176
9  0.51149425 0.00000000 0.4785295
10 1.00000000 0.23225806 0.2092856
11 0.17241379 0.14838710 0.1381467
12 0.24137931 1.00000000 1.0000000
```

CUADRO I.2. Práctica para la estandarización de datos

EJEMPLO. La Tabla I.3 muestra los datos morfológicos de varias especies de peces (archivo **Cuadro I.2.xls**). La Tabla I.4 muestra los datos estandarizados.

Tabla I.3. Variables morfológicas de diferentes especies de peces.

Especie	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
1	93,4	27,8	13,3	12,5	44,4
2	96,3	26,7	14,4	11,8	43,3
3	140,4	14,5	24,4	4,3	49,3
4	111	32,1	20,4	12,5	54,4
5	111	52,2	22,2	26,2	52,3
6	25	7,5	3	4,4	12,6
7	24	4,9	2,5	2,4	14,7
8	31	8,6	2,4	3,8	11,5
9	30	8	7	14	18
10	31	11,1	4,4	5,75	14,4

Tabla I.4. Valores estandarizados de los datos mostrados en la Tabla I.3.

Especie	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
1	0,596	0,484	0,495	0,424	0,767
2	0,621	0,461	0,545	0,395	0,741
3	1,000	0,203	1,000	0,080	0,881
4	0,747	0,575	0,818	0,424	1,000
5	0,747	1,000	0,900	1,000	0,951
6	0,009	0,055	0,027	0,084	0,026
7	0,000	0,000	0,005	0,000	0,075
8	0,060	0,078	0,000	0,059	0,000
9	0,052	0,066	0,209	0,487	0,152
10	0,060	0,131	0,091	0,141	0,068

Estadística descriptiva con R

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro I.2.R**, el cual es un archivo *script* que contiene las instrucciones para realizar la estadística descriptiva que se muestra en este ejemplo. Este archivo está en el CD que acompaña al libro en la carpeta Capítulo I. Los pasos a seguir son como se explicó en el Cuadro I.1.

Para asignar los ángulos a las variables, lo que se hace es dividir 360º por el doble del número de variables y a cada variable se le asignan dos ángulos, incluso a aquellas variables que no tengan valores negativos. Para cada variable, a los valores negativos se le asigna un valor de ángulo 180º mayor que el asignado a los valores positivos. Por ejemplo, supongamos que en la matriz de datos de la Tabla I.4 existiesen valores negativos. En ese caso el valor del ángulo de cada variable sería de 36º (360º/10). El ángulo de la variable 1 sería de 36º para los valores positivos y de 216º para los valores negativos, el de la variable 2 sería

72º para los valores positivos y 252º para los valores negativos, el de la variable 3 sería 108º para los valores positivos y 288º para los valores negativos, y así sucesivamente. La última variable tendría un ángulo de 180º para los valores negativos y de 360º para valores positivos. Es decir, en la última variable el proceso es a la inversa de lo que se hace con las otras variables.

I.1.3. REPRESENTACIÓN DE LAS COORDENADAS POLARES

Las coordenadas polares de cada elemento de la muestra o población se calculan determinando la resultante para cada uno de los ejes o variables, es decir, desplazando sucesivamente el punto en la dirección de cada eje una distancia igual al valor de la variable correspondiente: el ángulo resultante y la distancia al origen son las coordenadas polares. En general es más sencillo calcular y representar las correspondientes coordenadas rectangulares mediante la siguiente transformación:

$$X = \sum_{j=1}^{n} |(z_j)| cos(\alpha) \qquad Y = \sum_{j=1}^{n} |(z_j)| sen(\alpha)$$

donde *X* e *Y* son las coordenadas rectangulares para cada caso en el gráfico polar, *z* es el valor de ese caso para la variable *j*, α es el ángulo en grados asignado a la variable *j* y *n* es el número de variables. El valor de *n* siempre sería el doble del número de variables que realmente existen, como se mencionó anteriormente. En el Cuadro I.3 se explican cuales son los pasos a seguir.

Cuando se utiliza el programa EXCEL, es necesario pasar el ángulo a radianes y, por ello, las fórmulas a utilizar son las siguientes:

$$X = \sum_{j=1}^{n} |(z_j)| cos(\alpha \frac{\pi}{180}) \qquad Y = \sum_{j=1}^{n} |(z_j)| sen(\alpha \frac{\pi}{180})$$

Otro ejemplo de cálculo de coordenadas polares se puede ver en el Cuadro I.4. En este caso se realiza la representación polar de diferentes lagunas, diferenciándolas en base a sus características físico-químicas. Cuando existan varias réplicas, o como en este ejemplo en el que hay lagunas con varias estaciones, también es posible representar la media, en vez de los valores de cada una de las réplicas por separado.

Las coordenadas polares nos permiten ver de una forma gráfica las lagunas que tienen una composición más semejante para los parámetros que se han considerado, que en este caso eran solo 8 variables (temperatura, conductividad, pH, nitrato, nitrito, amonio, fosfato y silicato), pero podrían ser muchas más las variables a considerar.

CUADRO I.3. Cálculo de coordenadas polares

EJEMPLO. En la Tabla I.5 se exponen las puntuaciones obtenidas en un análisis discriminante aplicado al porcentaje de pigmentos de distintas especies marinas del fitoplancton. Para cada especie hay varias réplicas. Las funciones del análisis discriminante son variables que no necesitan ser estandarizadas a una escala de 0-1.

Tabla I.5. Puntuaciones de un análisis discriminante aplicado al porcentaje de pigmentos de distintas especies marinas de fitoplancton.

Especie	Función 1	Función 2	Función 3	Función 4	Función 5
1	-4,86	21,96	6,50	10,48	-6,92
1	-5,06	21,73	6,37	9,93	-6,80
1	-4,86	23,14	7,31	9,97	-6,93
2	-27,12	-53,23	2,36	-2,84	-2,81
2	-27,02	-54,25	-0,51	-4,56	-3,18
3	-24,54	-39,34	7,76	2,88	2,17
3	-25,25	-41,95	8,11	2,61	2,21
3	-25,31	-41,39	8,78	2,95	2,48
3	-25,00	-40,01	8,81	3,31	2,50
4	-8,95	24,19	13,55	-26,42	6,04
4	-8,85	23,97	12,44	-22,11	4,25
4	-9,38	24,85	13,58	-25,05	4,67
5	-4,70	19,20	4,24	11,32	-6,06
5	-4,82	19,41	4,39	11,54	-6,46
6	-0,97	14,67	-16,64	11,77	27,93
6	-1,39	13,07	-17,35	10,94	26,80
7	-2,93	-1,55	-28,86	-8,04	-7,80
7	-2,64	-0,86	-28,50	-7,71	-7,83
8	102,05	-20,51	6,71	0,19	1,02
8	104,50	-20,57	7,01	-0,88	0,35
9	2,43	-1,66	-39,35	-12,69	-13,38
9	0,67	-0,80	-42,21	-13,84	-14,76
9	4,46	-2,33	-41,16	-13,61	-14,30

El archivo **Cuadro I.3.xls** está diseñado para calcular las coordenadas polares de una matriz que contenga como máximo 1000 casos (filas) y 20 variables (columnas). Este archivo solo sirve para el caso en el que existan valores negativos.

Paso 1. Introducir los códigos de los casos en "DATOS" del archivo **Cuadro I.3.xls**.

Paso 2. Introducir los valores de cada variable para cada uno de los casos en "DATOS" del archivo **Cuadro I.3.xls**.

Paso 3. En "COORDENADAS EJE X" del archivo **Cuadro I.3.xls** aparecen las coordenadas *X* y en "COORDENADAS EJE *Y*" las coordenadas *Y* de todos los casos.

CUADRO I.3. *(Continuación)*

Paso 4. Cuando existan réplicas, en "MEDIAS COORDENADAS *X* e *Y*" del archivo **Cuadro I.3.xls** aparecen los valores medios, desviación típica y el error (desviación típica de la media) para cada uno de los grupos. Los resultados que se obtienen se muestran en la Tabla I.6.

Tabla I.6 Coordenadas polares de las funciones discriminantes de la Tabla I.5.

Especie	*X*	*Y*
1	-14,552	30,359
1	-14,180	29,577
1	-14,035	31,958
2	-39,630	-65,980
2	-37,958	-70,641
3	-34,573	-42,756
3	-35,792	-45,497
3	-35,897	-44,152
3	-35,494	-42,416
4	23,473	15,101
4	18,534	16,427
4	20,825	16,316
5	-14,405	26,183
5	-15,053	26,581
6	27,287	4,473
6	26,224	1,540
7	4,768	-35,369
7	4,819	-34,003
8	75,015	46,976
8	77,079	48,009
9	10,500	-45,029
9	9,775	-48,641
9	12,309	-46,741

Cálculo de coordenadas polares con R

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro I.3.R**, el cual es un archivo *script* que contiene las instrucciones para realizar el gráfico de coordenadas polares. Este archivo está en el CD que acompaña al libro en la carpeta Capítulo I. Para poder grabar los resultados es mejor pasar el archivo **Cuadro I.3.R** a un directorio del disco duro y seleccionar éste como directorio de trabajo.

CUADRO I.3. *(Continuación)*

Paso 1. Las instrucciones para realizar este tipo de importación de datos están en el Apéndice II en el apartado AII.4.2. En este ejemplo usaremos el archivo **Cuadro I.3.V.csv**, en el cual se encuentran los datos mostrados en la Tabla I.5.

Paso 2. La siguiente ventana muestra las instrucciones del archivo *script* **Cuadro I.3.R**. Lo único que hay que hacer es cambiar la sección modificable por el usuario, para representar las coordenadas polares de cualquier tipo de archivo de datos.

```
# Cuadro I.3
# Cálculo de coordenadas polares
# Gráfica correpondiente


#########################################################
# Sección modificable por el usuario
#########################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro I.3.V.csv",encoding="latin1")

# Selección de las variables de interés
varInteres<-c("Función.1","Función.2","Función.3","Función.4","Función.5")

# Selección de las variables de identificación
varID<-c("Especie")
# varID<-NULL

# Selección de la variable de agrupación
varAgrupa<-c("Especie")
# varAgrupa<-NULL

# Nombre del archivo de salida con coodenadas polares
nomCoorPol<-("Salida Cuadro I.3-1.V.csv")

# Nombre del archivo de salida con resumen de las coordenadas polares
# si se coloca una variable de agrupación
nomResumenCoordPol<-("Salida Cuadro I.3-2.V.csv")

# Título de la gráfica
grafTitulo<-"Gráfica de coordenadas polares"

# Título de la gráfica de los datos agrupados
grafTituloAg<-"Gráfica de coordenadas polares agrupada"

# Opción de ordenación de acuerdo al algoritmo propuesto
ordenar<-"No"
```

Paso 3. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro I.3.V.csv**. El comando encoding="latin1", permite importar fuentes como acentos y ñ en cualquier sistema operativo. Si lo que se representa es una salida de un análisis multivariante, como en este caso que son las puntuaciones de un discriminante, no es necesario que los datos estén en una escala de 0 a 1.

CUADRO I.3. *(Continuación)*

Sin embargo, si lo que se representa no es un conjunto de variables homogéneas o comparables (como la mayoría de las salidas de un análisis multivariante), es necesario estandarizar a una escala de 0 a 1 como se explica en los Cuadros I.1 o I.2.

```
datos<-read.csv2("Cuadro I.3.V.csv",encoding="latin1")
```

Paso 4. Se seleccionan las variables que se van a representar. El orden de las variables es importante porque, en el supuesto de que no se seleccione que el gráfico se realice por medio del algoritmo, las variables aparecerán en el gráfico en el orden que se especifique.

```
varInteres<-
c("Función.1","Función.2","Función.3","Función.4","Función.5")
```

Paso 5. Se seleccionan las variables que se usan como códigos de identificación, que pueden ser varias. A la hora de grabar los resultados en el primer archivo de salida, estas variables se utilizan para identificar cada una de las filas. En el caso de que no se quiera ningún código, se borra el símbolo # que está delante de varID<-NULL, permitiendo así que se ejecute ese comando y no se selecciona ninguna variable.

```
varID<-c("Especie")
# varID<-NULL
```

Paso 6. Si se elige alguna variable de agrupación, como es el caso en este ejemplo que se selecciona la variable «Especie», se genera un segundo archivo y un segundo gráfico con la media aritmética, desviación típica y el error, en aquellos casos que existan réplicas. Se pueden seleccionar varias variables de agrupación. En el caso de que no se quiera variable de agrupación, se borra el símbolo # que está delante de «varAgrupa<-NULL», permitiendo así que se ejecute ese comando. Es necesario para representar el gráfico seleccionar o variable de identificación o de agrupación.

```
varAgrupa<-c("Especie")
# varAgrupa<-NULL
```

Paso. 7. Es necesario definir los nombres de los archivos donde se grabarán las coordenadas polares.

```
nomCoorPol<-("Salida Cuadro I.3-1.V.csv")
```

y el archivo donde se grabará el resumen con los valores medios, desviaciones y errores. En el supuesto de que no existiera variable de agrupación, este archivo no se crearía.

```
nomResumenCoordPol<-("Salida Cuadro I.3-2.V.csv")
```

CUADRO I.3. *(Continuación)*

Paso 8. El siguiente paso es definir los nombres de los títulos de las dos gráficas que se generan. La primera gráfica representa los datos no agrupados.

grafTitulo<-"Gráfica de coordenadas polares"

La segunda gráfica solo se genera si se ha seleccionado alguna variable de agrupación.

grafTituloAg<-"Gráfica de coordenadas polares agrupada"

Paso 9. Por último, hay que especificar si se quiere que el gráfico de coordenadas polares se realice utilizando un algoritmo que tiende a buscar la máxima separación entre los distintos grupos, en cuyo caso se especifica "Si", o que se utilice el orden de las variables descritas en el paso 4, en cuyo caso se especifica "No". En este caso, como es una salida de un análisis multivariante, es mejor no usar el algoritmo y que la gráfica se realice en el mismo orden de las funciones discriminantes, es decir, el orden que se le asignó en el paso 4. Como norma general, siempre que se quiera representar las puntuaciones de un análisis multivariante es mejor no usar el algoritmo, ya que las funciones de salida de un análisis multivariado ya están ordenadas en función del peso de cada función.

ordenar<-"No"

Paso 10. Para ejecutar el archivo *script* se seleccionan todas las líneas del archivo con *ctrl-A* y luego se pulsa *ctrl-R*, para ejecutar todo el archivo a la vez.

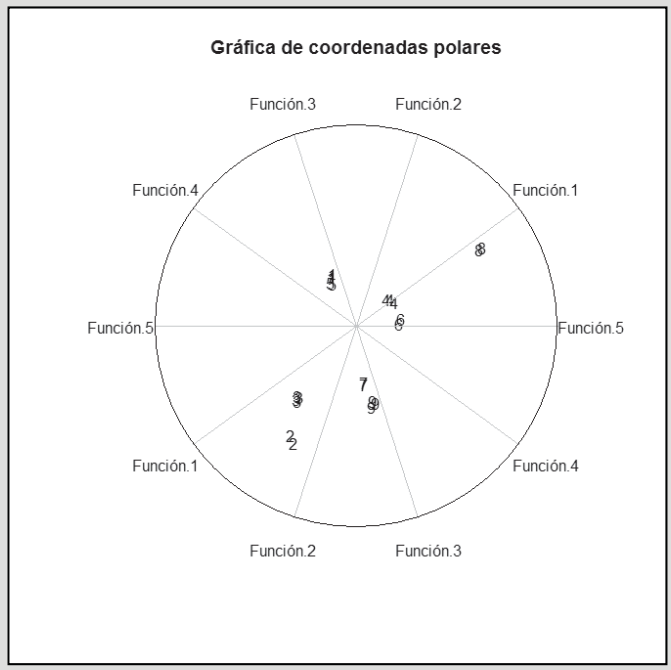
Paso 11. Aparecen dos ventanas con las dos gráficas. Para poder ver las gráficas es necesario arrastrar la ventana de la segunda gráfica, y así se puede ver la primera gráfica que está debajo.

La primera gráfica que se muestra a continuación representa los datos sin agrupar. La segunda gráfica muestra los promedios y desviaciones de los grupos especificados en la variable agrupación.

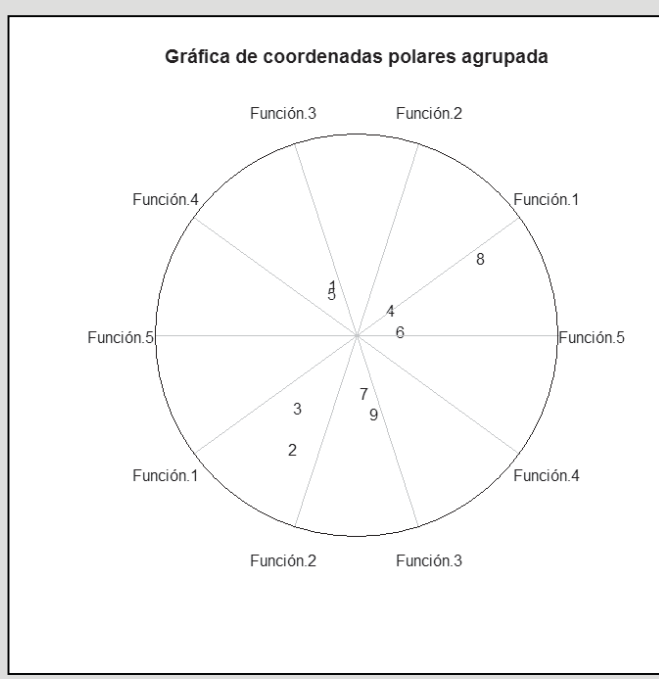
Es muy importante resaltar que no por poner más ejes en la gráfica (incluir muchas variables), la separación entre grupos será mayor. Si se ponen ejes que tengan poca carga en el analisis discriminante, o en cualquier otro tipo de analisis multivariante que se esté representando, esto puede provocar una mayor agrupación y, por tanto, un mayor solapamiento entre los grupos de datos y que el gráfico se vea menos claro. Por tanto, es importante representar solo aquellos ejes que tengan más carga o que sean significativos, para una mejor visualización del gráfico.

CUADRO I.3. *(Continuación)*

Gráfica que representa los valores sin agrupar:



Gráfica que representa el valor medio para el conjunto de datos:



CUADRO I.3. *(Continuación)*

Paso 12. En los archivos que se han creado **Salida Cuadro I.3-1.V** y **Salida Cuadro I.3-2.V**, es posible ver los datos usados para generar ambas gráficas de coordenadas polares.

	A	B	C	D
1	Especie	X	Y	
2	1	-14,552918	30,3704216	
3	1	-14,1806637	29,5872023	
4	1	-14,035983	31,9632536	
5	2	-39,6311873	-65,9902911	
6	2	-37,957095	-70,6421131	
7	3	-34,5679464	-42,7657933	
8	3	-35,7986042	-45,4912106	
9	3	-35,8862029	-44,1568312	
10	3	-35,4894808	-42,4220254	
11	4	23,4614677	15,1029086	
12	4	18,5405313	16,4301363	
13	4	20,8299178	16,3116557	
14	5	-14,397558	26,1839031	
15	5	-15,0540828	26,585062	
16	6	27,2984456	4,47449939	
17	6	26,2251174	1,54282727	
18	7	4,77333096	-35,3696329	
19	7	4,81294589	-34,0065967	
20	8	75,0150285	46,9705843	
21	8	77,0815222	48,0099815	
22	9	10,4991875	-45,0335044	
23	9	9,77523033	-48,6460725	
24	9	12,318067	-46,739683	
25				

	A	B	C	D	E	F	G
1	Especie	X.mean	Y.mean	X.sd	Y.sd	X.err	Y.err
2	1	-14,2565216	30,6402925	0,26668567	1,21079628	0,15397104	0,69905356
3	2	-38,7941412	-68,3162021	1,18376207	3,28933485	0,83704618	2,32591098
4	3	-35,4355586	-43,7089652	0,60292091	1,40508242	0,30146045	0,70254121
5	4	20,9439723	15,9482335	2,46245003	0,7344659	1,42169619	0,42404408
6	5	-14,7258204	26,3844825	0,46423311	0,28366216	0,32826238	0,20057944
7	6	26,7617815	3,00866333	0,75895762	2,07300523	0,53666408	1,46583606
8	7	4,79313842	-34,6881148	0,02801198	0,96381214	0,01980746	0,6815181
9	8	76,0482753	47,4902829	1,46123171	0,73496482	1,03324685	0,51969861
10	9	10,8641616	-46,80642	1,31011799	1,80720847	0,75639697	1,0433923
11							

CUADRO I.4. Práctica del cálculo de coordenadas polares

EJEMPLO. Para practicar el cálculo de coordenadas polares usaremos los datos físico-químicos estandarizados de las lagunas que se muestran en la Tabla I.2 (Cuadro I.1 y archivo **Cuadro I.1.xls**). En este caso, como los valores no son homogéneos, es necesario estandarizar los datos a una escala de 0-1. El archivo **Cuadro I.4.xls** está diseñado para calcular coordenadas polares en aquellos casos en los que no existan valores negativos y se trabaja con valores en una escala de 0-1. Los pasos a seguir son los mismos que los descritos anteriormente para el caso en el que sí había valores negativos (Cuadro I.3). Las coordenadas polares se muestran en la Tabla I.7.

Tabla I.7. Coordenadas polares de los valores estandarizados de los datos físico-químicos de las lagunas (Tabla I.2).

Laguna	Estación	*X*	*Y*
1	1	-0,546	-0,876
2	1	1,217	0,317
	2	1,528	-0,460
3	1	0,920	0,158
	2	0,663	-1,013
	3	0,680	-0,518
4	1	-0,417	-0,290
5	1	-0,032	-1,057
6	1	0,259	-0,290
7	1	-0,354	0,382
8	1	0,229	-0,979
9	1	0,958	-0,620

Coordenadas polares con R

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro I.4.R**, el cual es un archivo *script* que contiene las instrucciones para realizar el gráfico de coordenadas polares. Este archivo está en el CD que acompaña al libro en la carpeta Capítulo I.

Paso 1. En este caso se utiliza el archivo **Salida Cuadro I.1.V.csv** que es el archivo de salida del *script* **Cuadro I.1.R** en el que se estandarizaron los datos físico-químicos de las lagunas que se muestran en la Tabla I.2 (Cuadro I.1 y archivo **Cuadro I.1.xls**).

Paso 2. La siguiente ventana muestra las instrucciones del archivo *script* **Cuadro I.4.R**. Lo único que hay que hacer es cambiar la sección modificable por el usuario, para representar las coordenadas polares de cualquier archivo de datos.

CUADRO I.4. *(Continuación)*[R]

```
# Cuadro I.4
# Cálculo de coordenadas polares
# Gráfica correspondiente

#########################################################
# Sección modificable por el usuario
#########################################################
# Lectura de la base de datos
datos<-read.csv2("Salida Cuadro I.1.V.csv",encoding="latin1")

# Selección de las variables de interés
varInteres<-c("Temperatura","Conductividad","pH","Nitrato","Nitrito","Amonio","Fosfato","Silicato")

# Selección de las variables de identificación
# varID<-c("Laguna","Estación")
varID<-c("Laguna")
# varID<-NULL

# Selección de la variable de agrupación
# varAgrupa<-c("Laguna","Estación")
varAgrupa<-c("Laguna")
# varAgrupa<-NULL

# Nombre del archivo de salida con coodenadas polares
nomCoorPol<-("Salida Cuadro I.4-1.V.csv")

# Nombre del archivo de salida con resumen de las coordenadas polares
nomResumenCoordPol<-("Salida Cuadro I.4-2.V.csv")

# Título de la gráfica general
grafTitulo<-"Gráfica de coordenadas polares"

# Título de la gráfica de los datos agrupados
grafTituloAg<-"Gráfica de coordenadas polares agrupada"

# Opción de ordenación de acuerdo al algoritmo propuesto
ordenar<-"No"
```

Paso 3. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro I.1.V.csv**.

datos<-read.csv2("Salida Cuadro I.1.V.csv",encoding="latin1")

Paso 4. Se seleccionan las variables que se van a representar.

varInteres<-
c("Temperatura","Conductividad","pH","Nitrato","Nitrito","Amonio",
"Fosfato","Silicato")

Paso 5. Se seleccionan las variables que se usan como códigos de identificación, en este caso ninguna, porque se selecciona «varID<-NULL».

# varID<-c("Laguna","Estación")
# varID<-c("Laguna")
varID<-NULL

Paso 6. Si se elige alguna variable de agrupación, como en este ejemplo que se selecciona la variable "Laguna", se genera un segundo archivo y un segundo gráfico con la media aritmética, desviación típica y el error de la media, en aquellos casos en los que existan réplicas. Se pueden seleccionar varias variables de agrupación. Es necesario para representar el gráfico seleccionar o variable de identificación o de agrupación.

CUADRO I.4. *(Continuación)*[R]

```
# varAgrupa<-c("Laguna","Estación")
varAgrupa<-c("Laguna")
# varAgrupa<-NULL
```

Paso. 7. Es necesario definir los nombres de los archivos donde se grabarán las coordenadas polares

```
nomCoorPol<-("Salida Cuadro I.4-1.V.csv")
```

y el archivo donde se grabará el resumen con los valores medios, desviaciones y errores. En el supuesto de que no existiera variable de agrupación, este archivo no se crearía.

```
nomResumenCoordPol<-("Salida Cuadro I.4-2.V.csv")
```

Paso 8. El siguiente paso es definir los nombres de los títulos de las dos gráficas que se generan. La primera gráfica representa los datos no agrupados

```
grafTitulo<-"Gráfica de coordenadas polares"
```

la segunda gráfica solo se genera si se ha seleccionado alguna variable de agrupación

```
grafTituloAg<-"Gráfica de coordenadas polares agrupada"
```

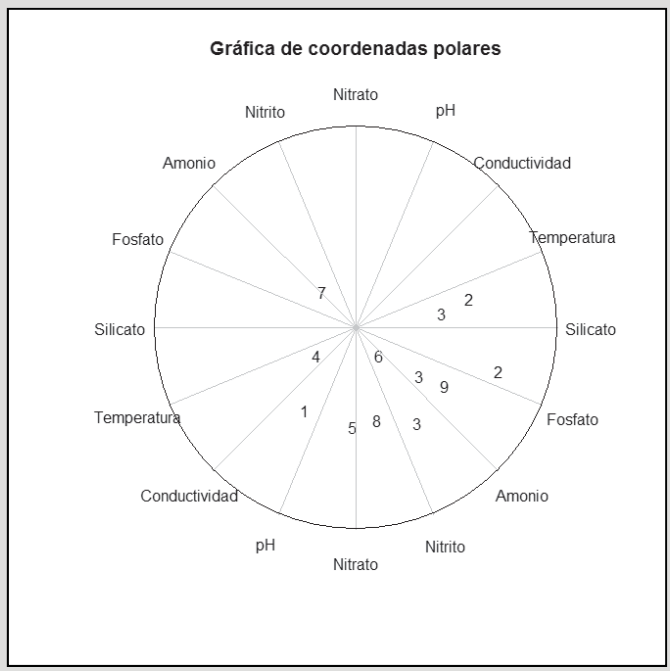
Paso 9. Por último, hay que especificar si se quiere que el gráfico de coordenadas polares se realice utilizando un algoritmo que tiende a buscar la máxima separación entre los distintos grupos, en cuyo caso se especifica "Si", o que se utilice el orden de las variables descritas en el paso 4, en cuyo caso se especifica "No". El algoritmo que se usa consiste en seleccionar, desde la variable con mayor desviación típica (primera variable) a la que tiene menor desviación típica (última variable). En este ejemplo, como no es una salida de un análisis multivariante, es mejor usar el algoritmo para que busque la mejor organización de las variables. Las gráficas siguientes muestran como es el resultado en ambos casos.

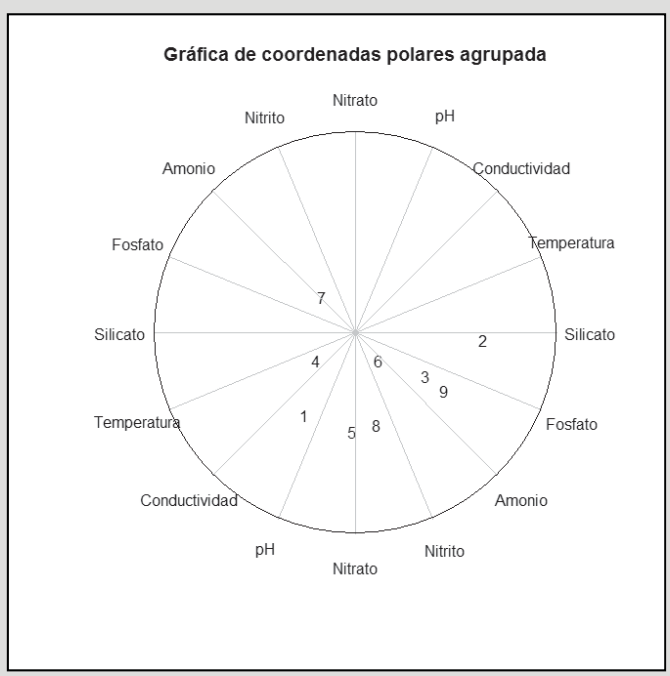
Paso 10. Para ejecutar el archivo *script* se seleccionan todas las líneas del archivo con *ctrl-A* y luego se pulsa *ctrl-R*, para ejecutar todo el archivo a la vez.

Paso 11. Aparecen dos ventanas con las dos gráficas. Para poder ver las gráficas es necesario arrastrar la ventana de la segunda gráfica, y así se puede ver la primera que está debajo. La primera gráfica representa los datos sin agrupar. La segunda gráfica, que se muestra a continuación, muestra los promedios y desviaciones de los grupos especificados en la variable agrupación.

CUADRO I.4. *(Continuación)*[R]

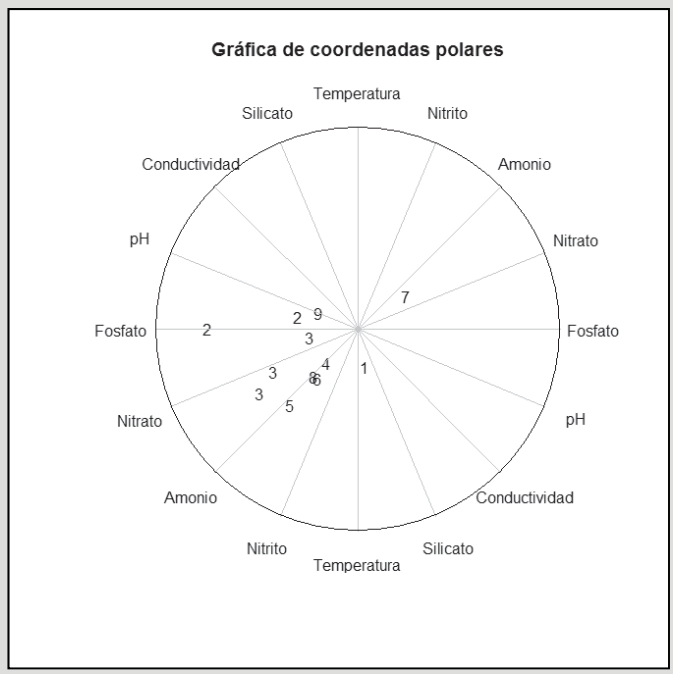
Gráficas de coordenadas polares de los valores no agrupados y agrupados, seleccionando la opción "No" y, por tanto, las variables salen en el orden especificado.

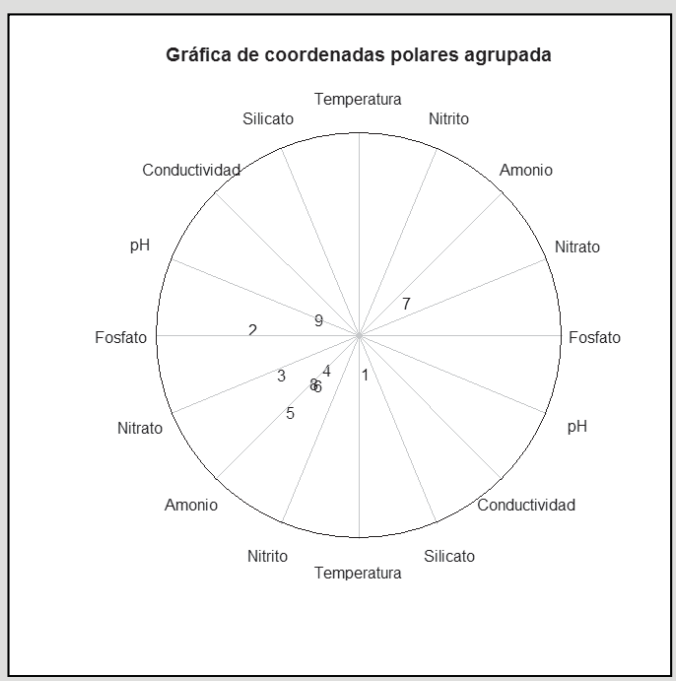




CUADRO I.4. *(Continuación)*[R]

Gráficas de coordenadas polares de los valores no agrupados y agrupados, seleccionando la opción "Si" y, por tanto, el gráfico de coordenadas polares se realiza utilizando un algoritmo que tiende a buscar la máxima separación entre los distintos grupos.





Capítulo II

ESTADÍSTICA DESCRIPTIVA

Dentro de la estadística descriptiva se incluyen medidas de posición como las medias aritmética, geométrica, armónica, ponderada, la moda y la mediana, y medidas de dispersión o variabilidad de la muestra como la varianza, la desviación típica y el coeficiente de variación.

El hecho de que sea una estadística simple y fácil de calcular no significa que no sea importante. Es necesario, por lo general, calcular algunas de las medidas citadas anteriormente para poder tener una idea preliminar de los datos antes de pasar a realizar un tratamiento más complejo de los mismos. Por ejemplo, el tipo de distribución que tengan los datos condicionará la clase de análisis estadístico que podremos utilizar. Algunos tratamientos estadísticos que se utilizan mucho, como por ejemplo el análisis de la varianza, se basan en la comparación de la media y dispersión de los datos, ambos incluidos dentro de la estadística descriptiva. Además, algunas medidas, como por ejemplo una variante de la media ponderada, pueden ser unos indicadores muy útiles por sí solos sin necesidad de tratamientos estadísticos más complejos.

II.1. Medidas de posición

El primer paso a la hora de tratar unos datos es intentar buscar algún tipo de medida que nos permita caracterizar, diferenciar y distinguir las series de datos. Esto se puede hacer determinando la posición de los datos. Dentro de este grupo se diferencian las llamadas medidas de posición central y otras medidas que estiman diferentes posiciones de los datos. Unas u otras se utilizan dependiendo del tipo de datos con el que se esté trabajando.

II.1.1. MEDIDAS DE POSICIÓN CENTRAL

II.1.1.1. Media aritmética

La media aritmética ($\bar{x}$), que también se conoce simplemente como media o promedio, se calcula por medio de la siguiente fórmula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

donde x son cada uno de los valores i de la variable y n el número de datos.

Cuando la media es del conjunto de toda la población se suele denotar por el símbolo μ y cuando es la media de una muestra de la población, entonces se utiliza $\bar{x}$.

En el caso de datos agrupados (tabla de frecuencias) la media se calcula por medio de la siguiente fórmula:

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{n}$$

donde m y f son el valor medio y la frecuencia (número de datos) de la clase i, respectivamente, y k el número de intervalos o clases.

En el archivo **Cuadro II.1.xls** se muestra un ejemplo de cómo se calcula la media aritmética de una muestra agrupando y sin agrupar los datos, el cual también está descrito en el Cuadro II.1.

La media aritmética es la que se utiliza más frecuentemente porque es la que tiene un error estándar más pequeño, es la más fácil de estimar, tiende hacia una distribución Normal incluso si los datos originales no presentan esta distribución y, por último, es más sensible que otras medidas de posición a cambios en la distribución de los datos, lo cual es muy importante en estadística para poder determinar diferencias entre series de datos (Sokal & Rohlf, 1981). El problema que plantea la media aritmética también deriva de su sensibilidad, ya que está más afectada por los datos raros que se salen del rango.

II.1.1.2. Media geométrica

La media geométrica (*MG*) se utiliza en algunos casos con datos relativos como porcentajes o tasas de incremento y se calcula por medio de la siguiente fórmula:

$$MG = \sqrt[n]{x_1 x_2 .... x_n}$$

La media geométrica es siempre inferior o igual a la media aritmética. En el archivo **Cuadro II.1.xls** y en el Cuadro II.1 se muestra un ejemplo para calcular la media geométrica.

II.1.1.3. Media armónica

La media armónica (*MA*) se utiliza, por ejemplo, en algunos casos en los que es necesario promediar variaciones con respecto al tiempo. Se calcula de la siguiente forma:

$$MA = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} ..... \frac{1}{x_n}\right)}$$

La media armónica es siempre inferior o igual a la media geométrica y, por tanto, también inferior o igual a la media aritmética. Al igual que las otras me-

dias descritas anteriormente, en el archivo **Cuadro II.1.xls** y en el Cuadro II.1 también se muestra un ejemplo para calcular la media armónica.

II.1.1.4. Moda

A veces es necesario determinar la posición de variables que no son numéricas sino cualitativas como, por ejemplo, las especies a las que pertenecen diferentes individuos. En este caso no podemos utilizar ninguna de las medias citadas anteriormente y, por lo tanto, debemos recurrir a medidas como la moda. También puede ser utilizada con variables cuantitativas.

La moda de un conjunto de observaciones se define como el valor que se presenta con mayor frecuencia, es decir, un mayor número de veces. La moda se puede denotar por Mo. Ejemplo en el archivo **Cuadro II.1.xls** y en el Cuadro II.1.

Si solamente es un valor el que se repite más veces, en ese caso se dice que la distribución es unimodal, porque solo tiene una moda. En el caso de que ningún valor se repita, entonces por definición se considera que no existe moda. Se puede dar el caso de que dos observaciones tengan la misma frecuencia máxima, siendo en este caso la distribución bimodal. La distribución multimodal sería en aquellas situaciones en las que existen más de dos valores con la misma frecuencia máxima. Por último, en situaciones raras en las que la distribución de los datos tenga forma de U, el punto medio de la distribución se denomina antimoda.

II.1.1.5. Mediana

La mediana de un conjunto de observaciones es el valor para el cual, cuando todas las observaciones se ordenan de manera creciente, la mitad de éstas son menores que este valor y la otra mitad son mayores. Sea x_1, x_2, ……x_n una muestra aleatoria de *n* observaciones ordenadas de forma creciente, la mediana de estos datos se calcula de la siguiente manera (Cuadro II.1, archivo **Cuadro II.1.xls**):

$$\text{Mediana} \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{si } n \text{ es un número impar} \\ \dfrac{x_{\frac{n}{2}} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{si } n \text{ es un número par} \end{cases}$$

La mediana se suele utilizar en series de datos que tienen una distribución muy asimétrica, ya que en estos casos no siempre es adecuado utilizar la media aritmética. Por último, no se ve afectada por valores raros que se salgan mucho del rango normal, como le ocurre a la media aritmética.

CUADRO II.1. Cálculo de las medias

X	22 23 25 22 25 25 24	26 26 29 27	32 34 34 33 35 33	36 36 39
Rango	20-25	26-30	31-35	36-40
f_i	7	4	6	3
m_i	22,5	28	33	38
f_a	7	11	17	20

1. Media aritmética para datos sin agrupar

$$\bar{x} = \frac{22+23+22+24+26+25+26+25+29+27+25+.....39}{20} = 29{,}3$$

2. Media aritmética agrupando los datos

$$\bar{x} = \frac{7*22{,}5+4*28+6*33+3*38}{20} = 29{,}08$$

3. Media geométrica

$$MG = \sqrt[20]{22*23*22*24*26*25*26*25*29*27*25*.....39} = 28{,}83$$

4. Media armónica

$$MA = \frac{1}{\frac{1}{20}\left(\frac{1}{22}+\frac{1}{23}+\frac{1}{22}+\frac{1}{24}+\frac{1}{26}+\frac{1}{25}+\frac{1}{26}+\frac{1}{25}+\frac{1}{29}+.....\frac{1}{39}\right)} = 28{,}37$$

5. Media acotada o recortada

En primer lugar se ordenan los datos como se muestra en la siguiente tabla

22
22
23
24
25
25
25
26
26
27
29
32
33
33
34
34
35
36
36
39

Media acotada al 5%, en otras palabras que de los 20 datos se recorta a cada lado el 5%, es decir un dato a cada lado, el máximo y el mínimo.

$$\bar{x} = \frac{23+24+25+25+25+26+26+27+29+.....36}{18} = 29{,}17$$

Media recortada al 10%. Eliminados del cálculo de la media los 2 datos más bajos y los 2 datos más altos.

$$\bar{x} = \frac{24+25+25+25+26+26+27+29+.....36}{16} = 29{,}19$$

En este caso se observa que la variación en la media recortada con respecto a la media original no es notable, por lo tanto los valores que están a los extremos de la distribución no son valores anómalos o que puedan influir negativamente en la media.

CUADRO II.1. *(Continuación)*

6. Moda

La moda es 25 porque es la observación que se repite más veces (3 veces).

7. Mediana

Para el cálculo de la mediana es necesario ordenar los datos como se muestran en la tabla del cuadro anterior. Como el número de datos es par (20 en este caso), entonces la mediana (*Mn*) se estima de la siguiente forma:

$$x_{\frac{n}{2}} = 27 \qquad x_{\frac{n}{2}+1} = 29 \qquad Mn = \frac{27+29}{2} = 28$$

II.1.1.6. Media acotada o recortada

Otra medida de tendencia central que se suele utilizar con fines exploratorios es la media acotada (en inglés*, trimmed mean*), porque es más resistente a valores extremos que la media y no es tan insensible como lo podría ser la mediana. La media acotada se define como el promedio de los valores en donde se ha excluido un porcentaje determinado de la población a cada lado de los valores extremos. Para su cálculo hay que seguir tres pasos:

1. Se ordenan los datos.
2. Se descarta el porcentaje de datos a cada extremo de la población o muestra.
3. Se calcula la media con los datos restantes.

Esta medida sirve como un parámetro exploratorio, ya que muestra el impacto de los valores extremos sobre la estimación de los promedios, porque se calcula a diferentes porcentajes. Además facilita la detección de valores inusuales, los cuales advierten al investigador de posibles errores, de datos que podrían no ser parte de la población objetivo del estudio o de valores anómalos que indiquen nuevos rumbos en la investigación.

II.1.1.7. Media ponderada

La media ponderada ($\bar{x}_w$) se utiliza en aquellos casos en que algunos datos son más fiables que otros, simplemente porque unos datos han sido calculados con un mayor esfuerzo o por cualquier otra razón y, por ello, es necesario darle más peso a estos valores a la hora de estimar la media. Se calcula por medio de la siguiente fórmula:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

donde *x* son cada uno de los valores *i* de la variable, *w* el peso que se le da al dato y *n* el número de datos.

Una aplicación de la media ponderada que merece una mención especial, porque puede dar una gran información, es la obtención del centro de gravedad de una distribución de datos de acuerdo a una determinada variable. En Ecología es un parámetro que permite cuantificar el nicho ecológico de una especie, ya que da información sobre el valor medio de su distribución para una determinada variable. Por ejemplo, se puede obtener la temperatura, la salinidad, concentración de oxígeno, etc. media en que suele aparecer una especie. En este caso, *w* son cada uno de los valores *i* de los datos de abundancia de la especie, *x* el valor de la variable para el valor *w* y *n* el número de datos. El cálculo de la media ponderada se explica en el Cuadro II.2. En el archivo **Cuadro II.2.xls** se muestra como se calcula con Excel.

CUADRO II.2. Cálculo de la media ponderada

El archivo de Excel donde se calcula la media ponderada es el **Cuadro II.2.xls** y los datos se muestran en la siguiente tabla:

Abundancia Especie 1 (*w*)	Abundancia Especie 2 (*w*)	Temperatura (*x*)
23	211	14,3
37	74	16,2
128	145	15,7
234	89	16,7
21	24	18,1
198	56	17,2
4	338	14,3
156	126	15,9

Media ponderada de la temperatura para la especie 1

$$T_1 = \frac{(23*14{,}3) + (37*16{,}2) + \ldots\ldots(156*15{,}9)}{23 + 37 + \ldots\ldots 156} = 16{,}44$$

Media ponderada de la temperatura para la especie 2

$$T_2 = \frac{(211*14{,}3) + (74*16{,}2) + \ldots\ldots(126*15{,}9)}{211 + 74 + \ldots\ldots 126} = 15{,}25$$

II.1.2. OTRAS MEDIDAS DE POSICIÓN

Se definen los cuantiles de orden *k* como los valores de la variable, supuesta ésta ordenada de menor a mayor, que la dividen en *k* partes con la misma frecuencia de observaciones. Por lo tanto existirán *k - 1* cuantiles de orden *k*.

El primer cuantil de orden *k* deja a su izquierda la fracción *1/k* de frecuencia de observaciones. El segundo cuantil de orden *k* deja a su izquierda la fracción *2/k* de frecuencia de observaciones. El r-ésimo cuantil de orden *k* deja a su izquierda la fracción r/k de frecuencia de observaciones. Por ejemplo, el cuantil 15 de orden 100 deja por debajo el 15% de los valores del total de la serie completa de valores.

Los cuantiles más utilizados son los **percentiles**, **cuartiles** y **deciles**, los cuales se describen a continuación.

Los **percentiles** son los 99 puntos que dividen la distribución en 100 partes, tales que dentro de cada una está incluido el 1% de los valores de la distribución.

Los **cuartiles** son los tres valores de la variable que dividen la distribución en 4 partes iguales, es decir, en 4 intervalos, dentro de cada cual está incluido el 25% de los valores de la distribución. El percentil 25 (P_{25}) sería igual al cuartil 1 (Q_1), el percentil 50 (P_{50}) sería igual al cuartil 2 (Q_2, igual también a la mediana), etc.

Los **deciles** son los 9 puntos que dividen la distribución en 10 partes, tales que dentro de cada una está incluido el 10% de los valores de la distribución. El percentil 10 (P_{10}) sería igual al decil 1 (D_1), el percentil 20 (P_{20}) sería igual al decil 2 (D_2), etc.

El procedimiento para encontrar el valor del cuantil *j* de orden *k* a partir de datos ordenados de menor a mayor, es el siguiente (ejemplo en el Cuadro II.3 y archivo **Cuadro II.1.xls**):

1. Encontrar la posición *i* del cuantil *j* -ésimo mediante el cálculo de *nj/k*.
2. Si *nj/k* no es un entero, entonces la posición *i* es el siguiente entero más grande y el valor del cuantil es el dato ordenado en la posición de este entero más grande.
3. Si *nj/k* es un entero, entonces la posición del cuantil será $i = nj/k + 0{,}5$ y así el valor del cuantil es el promedio de las observaciones ordenadas *nj/k* y *nj/k* + 1.

CUADRO II.3. Cálculo de los cuantiles

Vamos a utilizar la serie de datos del Cuadro II.1.

1. Percentil 25 (P_{25}) y Cuartil 1 (Q_1)

$$\frac{nj}{k} = \frac{20*25}{100} = 5$$

Al ser entero el valor será:

$$P_{25} = \frac{X_5 + X_6}{2} = \frac{25+25}{2} = 25$$

2. Percentil 75 (P_{75}) y Cuartil 3 (Q_3)

$$\frac{nj}{k} = \frac{20*75}{100} = 15$$

Al ser entero el valor será:

$$P_{75} = \frac{X_{15} + X_{16}}{2} = \frac{34+34}{2} = 34$$

3. Percentil 10 (P_{10}) y Decil 1 (D_1)

$$\frac{nj}{k} = \frac{20*10}{100} = 2$$

Al ser entero el valor será:

$$P_{10} = \frac{X_2 + X_3}{2} = \frac{22+23}{2} = 22{,}5$$

4. Percentil 16 (P_{16})

$$\frac{nj}{k} = \frac{20*16}{100} = 3{,}2$$

Al no ser entero, el valor es el dato ordenado en la posición del siguiente entero más grande:

$$P_{16} = X_4 = 24$$

II.1. Medidas de dispersión

Además de la posición, es también importante la dispersión o variabilidad de los datos. La Figura II.1 muestra dos series de datos que tienen la misma media (10,55 ºC), pero la variabilidad de la temperatura es mayor en un hábitat que en el otro.

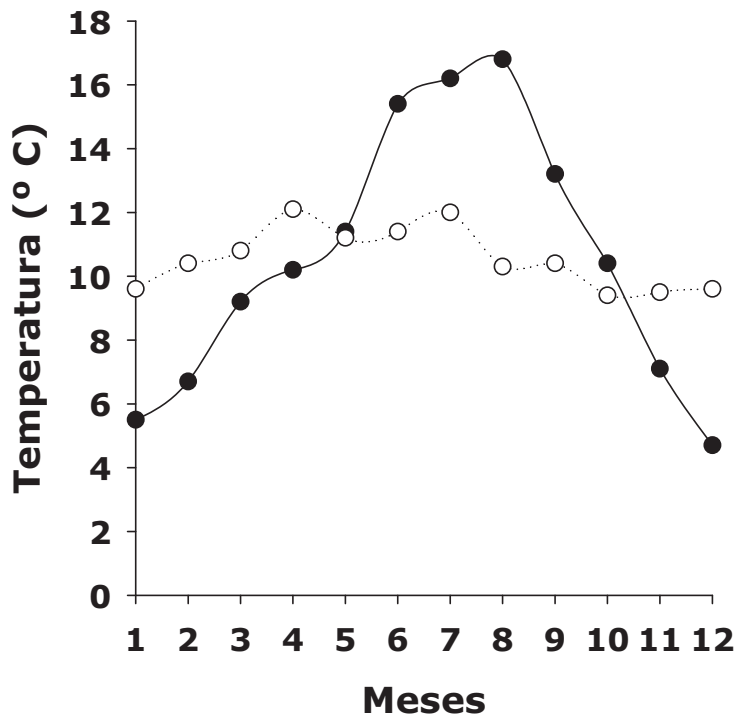


Figura II.1. Temperatura a lo largo del año en dos hábitats.

Las medidas de dispersión tienen como finalidad estudiar hasta qué punto, para una determinada distribución de datos, las medidas de posición representan bien el conjunto de datos de la distribución.

Por ejemplo, si se quiere determinar si una media aritmética nos marca una tendencia central generalizable del comportamiento de todos los elementos del conjunto estudiado, tendremos que fijarnos en la separación o desviación de cada valor respecto a la media. Si todos los valores están cercanos al valor medio, éste será representativo de ellos. Es decir, será más representativa la media aritmética de una variable cuanto más agrupados en torno a ella estén los valores promediados y, por el contrario, será tanto más rechazable, por no ser representativa, cuanta mayor dispersión exista de los valores de la variable respecto a la media.

Por lo tanto, para complementar la información que se obtiene a partir de la media es necesario otro tipo de parámetros que midan la dispersión o variabilidad de los datos, los cuales se muestran a continuación.

II.2.1. AMPLITUD

El método más simple de estimar la dispersión de los datos es por medio de la amplitud, también conocida como rango, es decir, la diferencia entre el valor mínimo y el máximo de la serie de datos (Cuadro II.4 y archivo **Cuadro II.4.xls**).

II.2.2. VARIANZA Y CUASIVARIANZA

La mejor forma de medir la dispersión de una serie de datos es comparar cada uno de ellos con la media de la serie, y esto es exactamente lo que hace la varianza (σ^2) (Cuadro II.4 y archivo **Cuadro II.4.xls**):

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

A menudo queremos utilizar la varianza de una muestra como valor aproximado o estimado de la varianza desconocida de la población de la que procede esa muestra. En estos casos, el error cometido es generalmente más pequeño si en vez de considerar como estimador la varianza de la muestra, se utiliza lo que se denomina cuasivarianza (s^2), que se calcula como la anterior, pero cambiando el denominador por n - 1 (Cuadro II.4 y archivo **Cuadro II.4.xls**):

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

En aquellos casos en los que disponemos de toda la población se debe emplear n y no (n - 1). Sin embargo, como en general nuestros datos son una muestra de una población mayor, lo correcto es usar la cuasivarianza en vez de la varianza.

Muchas veces se hace referencia a la varianza cuando en realidad se está calculando la cuasivarianza (varianza estimada). La mayoría de los programas de estadística utilizan solo la cuasivarianza y no la varianza. Es frecuente también observar que σ^2 y s^2 se usan indistintamente para indicar varianza o cuasivarianza sin un criterio definido.

II.2.3. DESVIACIÓN TÍPICA Y CUASIDESVIACIÓN TÍPICA

El problema que presenta la varianza es que, para evitar valores negativos, las diferencias se elevan al cuadrado. Para poder tener la medida de dispersión en las mismas unidades que la media, se suele utilizar, en vez de la varianza, la desviación típica (σ), también llamada desviación estándar, y que se calcula simplemente como la raíz cuadrada de la varianza (Cuadro II.4 y archivo **Cuadro II.4.xls**).

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}$$

Al igual que ocurría con la varianza, salvo en el caso de que el tamaño de la muestra sea idéntico al de la población, es preferible utilizar la cuasidesviación típica (*s*), en vez de la desviación típica. Para calcular la cuasidesviación típica se divide por el número de grados de libertad (*n* - 1) en vez de dividir por el total de datos (*n*) (Cuadro II.4 y archivo **Cuadro II.4.xls**) como se muestra en la siguiente fórmula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

Es necesario mencionar que la mayoría de los programas estadísticos solo utilizan la cuasidesviación típica y, al igual que ocurría con la varianza, es frecuente observar que se hace mención a la desviación típica cuando en realidad se ha calculado la cuasidesviación típica. También se observa con frecuencia que σ y *s* se usan indistintamente para desviación o cuasidesviación típica sin un criterio definido.

II.2.4. DESVIACIÓN ABSOLUTA RESPECTO A LA MEDIA Y LA MEDIANA

Una medida de dispersión que se suele utilizar es la desviación media absoluta que se define en el caso de la media como:

$$DAM = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

y para la mediana como:

$$DAMed = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - Mediana\right|$$

Esta medida indica el promedio de las desviaciones ya sean estas negativas o positivas con respecto al promedio general y facilita la interpretación del grado de dispersión de los datos analizados.

CUADRO II.4. Cálculo de las medidas de dispersión

Como ejemplo vamos a utilizar la misma tabla de datos del Cuadro II.1.

1. Amplitud: 39- 22 =17

2. Varianza

$$\sigma^2 = \frac{(22-29{,}3)^2 + (23-29{,}3)^2 + (22-29{,}3)^2 + \ldots\ldots\ldots (39-29{,}3)^2}{20} = 27{,}61$$

3. Cuasivarianza

$$s^2 = \frac{(22-29{,}3)^2 + (23-29{,}3)^2 + (22-29{,}3)^2 + \ldots\ldots\ldots (39-29{,}3)^2}{19} = 29{,}06$$

4. Desviación típica

$$\sigma = \sqrt{\frac{(22-29{,}3)^2 + (23-29{,}3)^2 + (22-29{,}3)^2 + \ldots\ldots\ldots (39-29{,}3)^2}{20}} = 5{,}25$$

5. Cuasidesviación típica

$$s = \sqrt{\frac{(22-29{,}3)^2 + (23-29{,}3)^2 + (22-29{,}3)^2 + \ldots\ldots\ldots (39-29{,}3)^2}{19}} = 5{,}39$$

6. Desviación absoluta respecto de la media

$$DAM = \frac{|22-29{,}3| + |23-29{,}3| + |22-29{,}3| + \ldots\ldots\ldots |39-29{,}3|}{20} = 4{,}83$$

7. Desviación absoluta respecto de la mediana

$$DAMed = \frac{|22-28| + |23-28| + |22-28| + \ldots\ldots\ldots |39-28|}{20} = 4{,}8$$

8. Coeficiente de variación

$$CV = \frac{5{,}39 * 100}{29{,}3} = 18{,}4\%$$

9. Error estándar de la media

$$ES = \frac{5{,}39}{\sqrt{20}} = 1{,}21$$

10. Recorrido intercuartílico

$$Q = P_{75} - P_{25} = 34 - 25 = 9$$

II.2.5. COEFICIENTE DE VARIACIÓN

Cuando se comparan series de datos que tienen medias diferentes, la desviación típica no permite comparar adecuadamente cual de las dos series de datos presenta una mayor variación, ya que normalmente la serie de datos con una mayor media también suele tener una mayor variabilidad. En este caso es más aconsejable el uso del coeficiente de variación (*CV*), que se calcula como el porcentaje que representa la desviación típica con respecto a la media. En el caso de que la desviación típica sea desconocida porque nuestros datos son de una muestra, se utiliza la cuasidesviación típica como estimador de la desviación típica poblacional (ejemplo en la continuación del Cuadro II.4 o archivo **Cuadro II.4.xls**):

$$CV = \frac{s * 100}{\bar{x}}$$

II.2.6. ERROR ESTÁNDAR DE LA MEDIA

Proporciona una medida de la precisión de la estimación de la media poblacional a partir de una muestra, mientras que la desviación típica mide la variabilidad de los datos respecto de la media en la muestra. El error estándar se calcula a partir de la desviación típica. Cuando ésta es desconocida, se utiliza la cuasidesviación típica para obtener el error estándar estimado (ejemplo en la continuación del Cuadro II.4 o archivo **Cuadro II.4.xls**):

$$ES = \frac{s}{\sqrt{n}}$$

II.2.7. RECORRIDO INTERCUARTÍLICO

El recorrido intercuartílico (*Q*) se calcula a partir de los percentiles 75 (P_{75}) y 25 (P_{25}) de la siguiente forma (ejemplo en la continuación del Cuadro II.4 o archivo **Cuadro II.4.xls**):

$$Q = P_{75} - P_{25} = Q_3 - Q_1$$

Esta medida de dispersión se utiliza cuando se expresa la posición central por medio de la mediana.

II.3. Estadística descriptiva con R

En el Cuadro II.5 se explican los pasos a seguir para usar un *script* diseñado para R que permite calcular todas las medidas de posición y dispersión descritas anteriormente.

CUADRO II.5. Estadística con R

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro II.5.R**, el cual es un archivo *script* que contiene las instrucciones para realizar todas las medidas descritas anteriormente. Este archivo está en el CD que acompaña al libro en la carpeta Capítulo II. Para facilitar la grabación de los resultados es conveniente pasar el archivo **Cuadro II.5.R** a un directorio del disco duro y seleccionarlo como directorio de trabajo.

Para que este *script* funcione correctamente es necesario instalar los paquetes "psych", "class" y "e1071" como se explica en el Apéndice II.

Paso 1. La siguiente ventana muestra las instrucciones del archivo *script* **Cuadro II.5.R**. Lo único que hay que hacer es cambiar la sección modificable por el usuario.

```
# Cálculo de los estadísticos básicos

#########################################################
# Sección modificable por el usuario
#########################################################

# Lectura de la base de datos.
datos<-read.csv2("Cuadro II.5.V.csv",enc="latin1")

# Selecciòn de variables a trabajar

variables<-c("M1","M2","M6","M12")
#variables<-c("M1","M2","M3","M4","M5","M6","M7","M8","M9","M10","M11","M12","M13","M14",
"M15","M16","M19","M20","M21","M22","M23","M24","M25","M26","M27","M30","M31","M32")

# Selección de la o las variables de agrupación
#agrupa<-NULL
agrupa<-c("Especie")
#agrupa<-c("Orden","Familia")
#agrupa<-c("Orden","Familia","Género")

# Grado de recorte de la media acotada
recorte<-0.05

# Vector de cuantiles que se quieren calcular
cuantiles<-c(0.025,0.05,0.95,0.975)

# Nombre del archivo de salida
nombreArchivoSalida<-"Salida Cuadro II.5.V.csv"

# Seleccion de las variables para obtener el promedio ponderado
# Si es nula no se realiza el procedimiento de la media ponderada
#variablespondera<-NULL
#variablespondera<-c("M2")
variablespondera<-c("M6","M12")

# Selección de la variables de ponderación
# Si es nula no se realiza el procedimiento de la media ponderada
#variablepeso<-NULL
variablepeso<-c("M1")
#variablepeso<-c("M2","M12")

# Selección de la o las variables de agrupación por la cual se quiere obtener la ponderación
#agrupapondera<-NULL
agrupapondera<-c("Especie")
#agrupapondera<-c("Orden","Familia")

# Nombre del archivo de salida para los promedios ponderados
nombreArchivoSalidaPondera<-"Salida Cuadro II.5.V.Media ponderada.csv"
```

CUADRO II.5. *(Continuación)*[R]

Paso 2. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro II.5.V.csv** que usaremos a modo de ejemplo. El archivo contiene medidas morfométricas de distintas especies de peces.

```
datos<-read.csv2("Cuadro II.5.V.csv",enc="latin1")
```

Paso 3. Se seleccionan las variables de las que se quiere calcular las medidas estadísticas. En el ejemplo se seleccionan solo las variables M1, M2, M6 Y M12, ya que no tiene el símbolo # al inicio, pero se pone otro caso en el que se seleccionan todas las variables, aunque como vemos no se usaría ya que tiene el símbolo # al inicio de la línea.

```
variables<-c("M1","M2","M6","M12")
#variables<-
c("M1","M2","M3","M4","M5","M6","M7","M8","M9","M10","M11","M12","M
13","M14","M15","M16","M19","M20","M21","M22","M23","M24","M25","M
26","M27","M30","M31","M32")
```

Paso 4. Es posible hacer los cálculos estadísticos agrupando las variables de la forma que se quiera. En el caso de seleccionar «NULL», no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna. En el ejemplo se selecciona que los cálculos se hagan a nivel de especie.

```
#agrupa<-NULL
agrupa<-c("Especie")
#agrupa<-c("Orden","Familia")
#agrupa<-c("Orden","Familia","Género")
```

Paso 5. En el caso de estar interesados en la media acotada y en el cálculo de algún cuantil es necesario especificar el punto de corte de la media y los valores de los cuantiles que se quieren calcular.

```
recorte<-0.05

cuantiles<-c(0.025,0.05,0.95,0.975)
```

Paso 6. Es necesario especificar el nombre del archivo donde se guardarán todos los cálculos realizados.

```
nombreArchivoSalida<-"Salida Cuadro II.5.V.csv"
```

Paso 7. Si interesa la media ponderada, hay que especificar en primer lugar cual es la variable o variables para las que se va obtener el promedio ponderado. Si se selecciona «NULL» entonces no se calcula la media ponderada. En este caso se seleccionan las variables M6 y M12.

```
#variablespondera<-NULL
#variablespondera<-c("M2")
variablespondera<-c("M6","M12")
```

CUADRO II.5. *(Continuación)*[R]

Paso 8. A continuación hay que seleccionar la variable o variables que se usan para ponderar. Al igual que antes, si se selecciona «NULL» entonces no se calcula la media ponderada. Se selecciona la variable M1.

```
#variablepeso<-NULL
variablepeso<-c("M1")
#variablepeso<-c("M2","M12")
```

Paso 9. En el supuesto que se quiera realizar agrupaciones, se especifica a continuación. En este caso se selecciona que se realicen los cálculos a nivel de especie.

```
#agrupapondera<-NULL
agrupapondera<-c("Especie")
#agrupapondera<-c("Orden","Familia")
```

Paso 10. Por último, se especifica el nombre del archivo de salida donde se guardan los cálculos realizados de media ponderada.

```
nombreArchivoSalidaPondera<-"Salida Cuadro II.5.V. Media ponderada.csv"
```

Paso 11. Los datos que se muestran a continuación son los archivos con los resultados: **Salida Cuadro II.5.V.csv** y **Salida Cuadro II.5.V.Media ponderada.csv**.

	A	B	C	D	E	F	G	H
1	variables	IDgrupo	N	noNA	minimo	maximo	media	mediaAcotada
2	M1	Archosargus rhomboidalis	1	1	194,11	194,11	194,11	194,11
3	M2	Archosargus rhomboidalis	1	1	21,12	21,12	21,12	21,12
4	M6	Archosargus rhomboidalis	1	1	81,93	81,93	81,93	81,93
5	M12	Archosargus rhomboidalis	1	1	112,52	112,52	112,52	112,52
6	M1	Caranx latus	2	2	485,78	485,89	485,835	485,835
7	M2	Caranx latus	2	2	37,97	39,12	38,545	38,545
8	M6	Caranx latus	2	2	185,99	188,03	187,01	187,01
9	M12	Caranx latus	2	2	253,02	256,12	254,57	254,57
10	M1	Cetengraulis edentulus	1	1	127,79	127,79	127,79	127,79
11	M2	Cetengraulis edentulus	1	1	8,9	8,9	8,9	8,9

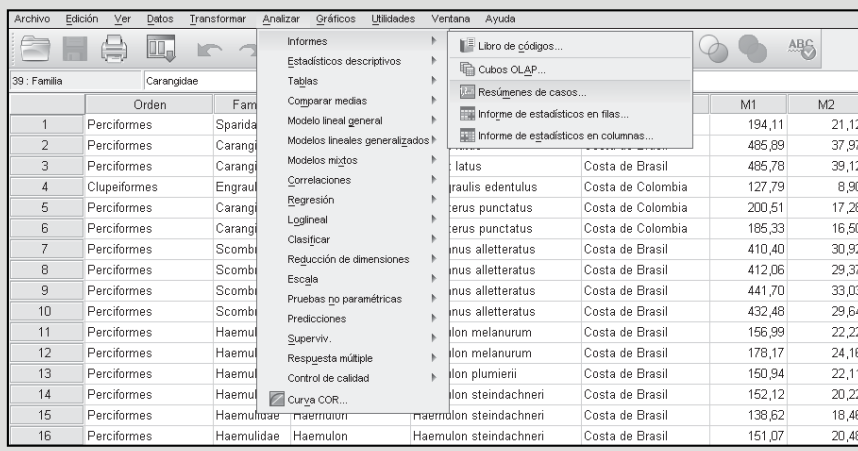
	A	B	C
1	IDgrupos	M6	M12
2	Archosargus rhomboidalis	81,93	112,52
3	Caranx latus	187,010115	254,570175
4	Cetengraulis edentulus	70,09	20,9
5	Decapterus punctatus	70,2338521	113,575339
6	Euthynnus alletteratus	126,728636	277,928697
7	Haemulon melanurum	69,2184193	92,6639232
8	Haemulon plumierii	61,2	78,37
9	Haemulon steindachneri	57,0493807	75,5311632
10	Lagocephalus laevigatus	216,976	32,504

II.4. Estadística descriptiva con SPSS

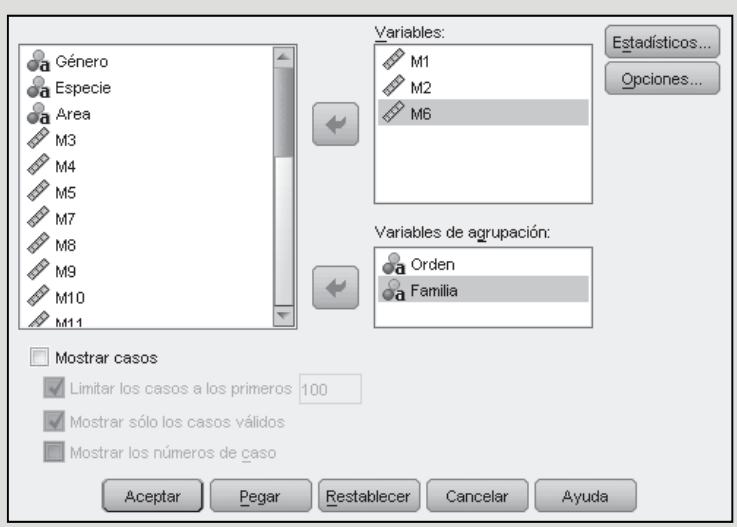
El programa SPSS permite calcular de forma sencilla las medidas de posición y dispersión descritas anteriormente. El proceso se describe en el Cuadro II.6.

CUADRO II.6. Estadística descriptiva con el programa SPSS

Paso 1. Se utilizan los mismos datos que en el ejemplo con R, los cuales están en el archivo **Cuadro II.6.sav**. Se entra en «Analizar» ⇨ «Informes» ⇨ «Resúmenes de casos…»

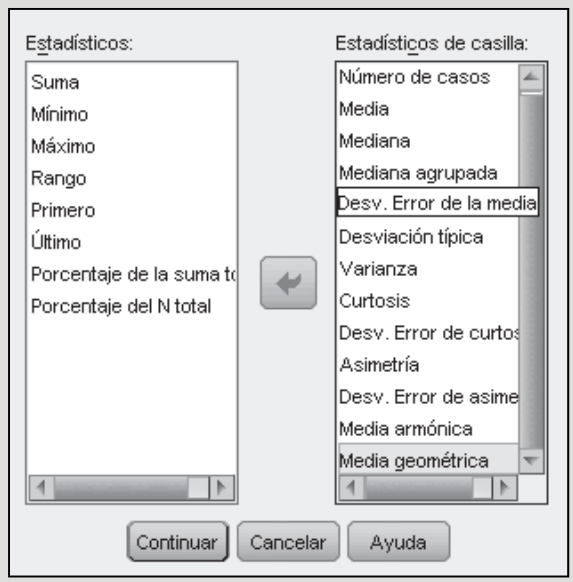


Paso 2. En la siguiente ventana se especifican las variables cuyos estadísticos queremos calcular. En el caso de querer realizar los cálculos agrupando los datos, se introduce en «Variables de agrupación» las variables que usamos para agrupar. Es conveniente no marcar «Mostrar casos», para que en la salida no aparezca todo el listado de datos.



CUADRO II.6. *(Continuación)*[SPSS]

Paso 3. Si se pulsa en la ventana anterior en «Estadísticos...» aparece la siguiente ventana donde se especifican las medidas que se quiere calcular. Para ello, simplemente se pasan a la ventana de la derecha los estadísticos seleccionados de la ventana de la izquierda.



Paso 4. La siguiente tabla muestra los resultados que se obtienen.

Resúmenes de casos

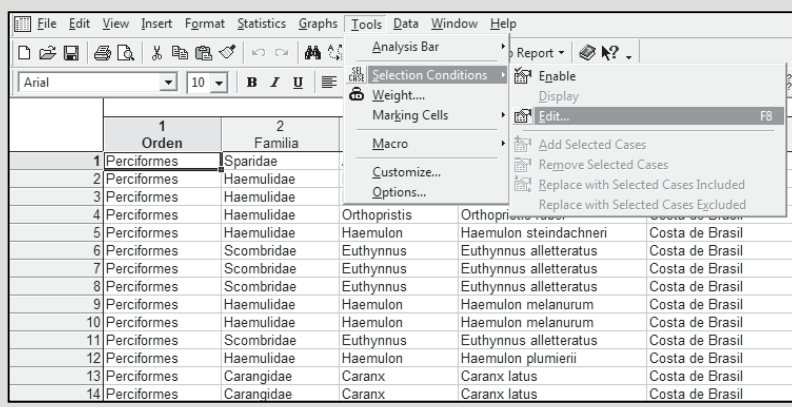
Orden	Familia		M1	M2	M6
Clupeiformes	Clupeidae	N	5	5	5
		Media	189,0380	13,0940	80,8100
		Mediana	185,3900	12,4600	77,0300
		Mediana agrupada	185,3900	12,4600	77,0300
		Error típ. de la media	3,83254	,55263	2,49694
		Desv. típ.	8,56982	1,23571	5,58333
		Varianza	73,442	1,527	31,174
		Curtosis	-3,129	-3,220	-3,318
		Error típ. de la curtosis	2,000	2,000	2,000
		Asimetría	,456	,530	,600
		Error típ. de la asimetría	,913	,913	,913
		Media armónica	188,7309	13,0034	80,5089
		Media geométrica	188,8838	13,0482	80,6581
	Engraulidae	N	1	1	1
		Media	127,7900	8,9000	70,0900
		Mediana	127,7900	8,9000	70,0900
		Mediana agrupada	127,7900	8,9000	70,0900
		Error típ. de la media	.	.	.
		Desv. típ.	.	.	.

II.5. Estadística descriptiva con STATISTICA

El programa STATISTICA también permite calcular la mayoría de las medidas de posición y dispersión descritas anteriormente. El proceso se describe en el Cuadro II.7.

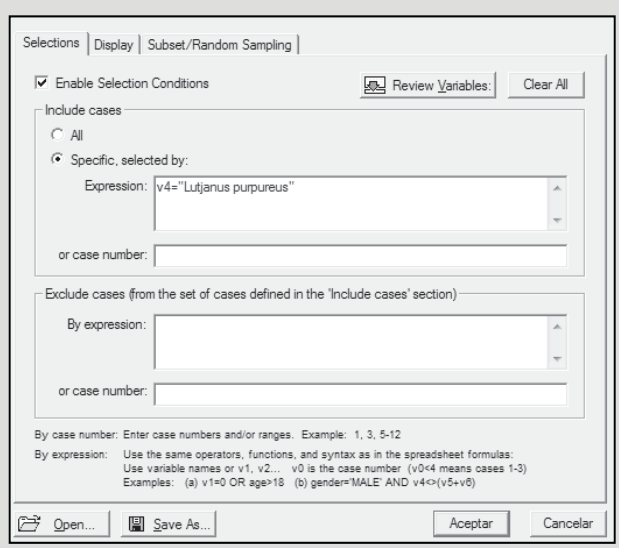
CUADRO II.7. Estadística descriptiva con el programa STATISTICA

Paso 1. Se utilizan los mismos datos que en el ejemplo con R, los cuales están en el archivo **Cuadro II.7.sta**. En el caso de que se quiera hacer el análisis de un determinado grupo de los datos se selecciona «Tools» ⇨ «Selection Conditions» ⇨ «Edit…».



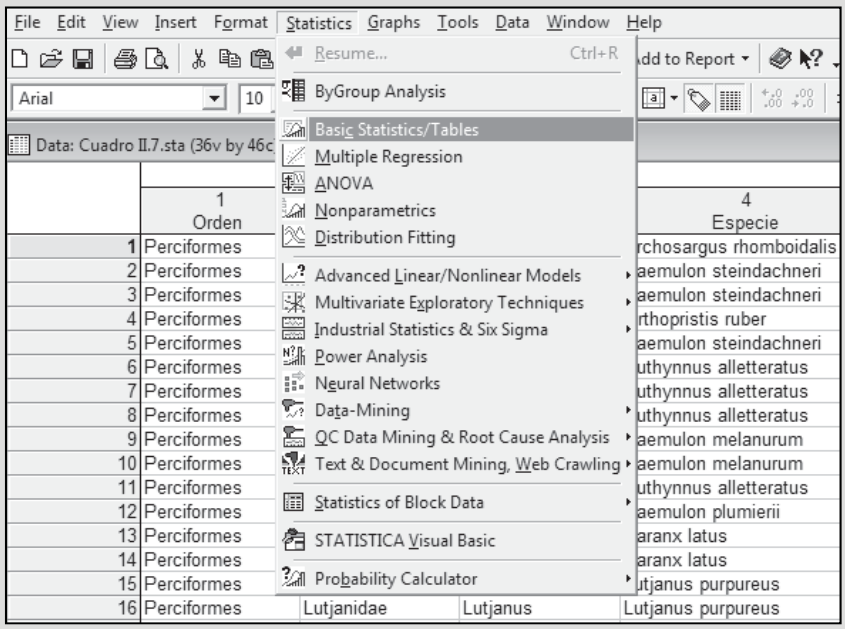
	1 Orden	2 Familia			
1	Perciformes	Sparidae			
2	Perciformes	Haemulidae			
3	Perciformes	Haemulidae			
4	Perciformes	Haemulidae	Orthopristis	Orthopristis	
5	Perciformes	Haemulidae	Haemulon	Haemulon steindachneri	Costa de Brasil
6	Perciformes	Scombridae	Euthynnus	Euthynnus alletteratus	Costa de Brasil
7	Perciformes	Scombridae	Euthynnus	Euthynnus alletteratus	Costa de Brasil
8	Perciformes	Scombridae	Euthynnus	Euthynnus alletteratus	Costa de Brasil
9	Perciformes	Haemulidae	Haemulon	Haemulon melanurum	Costa de Brasil
10	Perciformes	Haemulidae	Haemulon	Haemulon melanurum	Costa de Brasil
11	Perciformes	Scombridae	Euthynnus	Euthynnus alletteratus	Costa de Brasil
12	Perciformes	Haemulidae	Haemulon	Haemulon plumierii	Costa de Brasil
13	Perciformes	Carangidae	Caranx	Caranx latus	Costa de Brasil
14	Perciformes	Carangidae	Caranx	Caranx latus	Costa de Brasil

Paso 2. En la pestaña «Selections» se marca «Enable Selection Conditions» ⇨ «Specific, selected by:» donde se selecciona la variable o variables a analizar, usando OR y AND si se quieren combinar varias variables. En este caso se selecciona «v4="Lutjanus purpureus"».

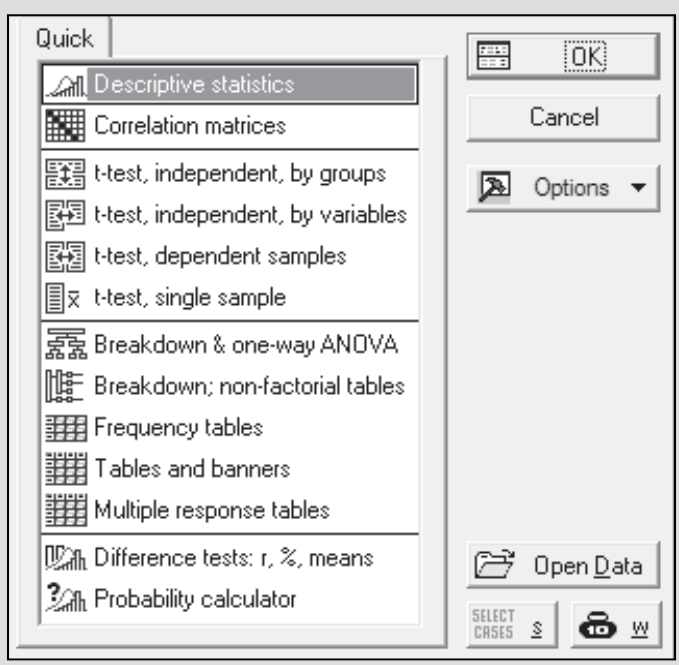

Selections
Display
Subset/Random Sampling
Enable Selection Conditions
Review Variables:
Clear All
Include cases
All
Specific, selected by:
Expression: v4="Lutjanus purpureus"
or case number:
Exclude cases (from the set of cases defined in the 'Include cases' section)
By expression:
or case number:
By case number: Enter case numbers and/or ranges. Example: 1, 3, 5-12
By expression: Use the same operators, functions, and syntax as in the spreadsheet formulas:
Use variable names or v1, v2... v0 is the case number (v0<4 means cases 1-3)
Examples: (a) v1=0 OR age>18 (b) gender='MALE' AND v4<>(v5+v6)
Open...
Save As...
Aceptar
Cancelar

CUADRO II.7. *(Continuación)* STATISTICA

Paso 3. Posteriormente se selecciona «Statistics» ⇨ «Basic Statistics/Tables».

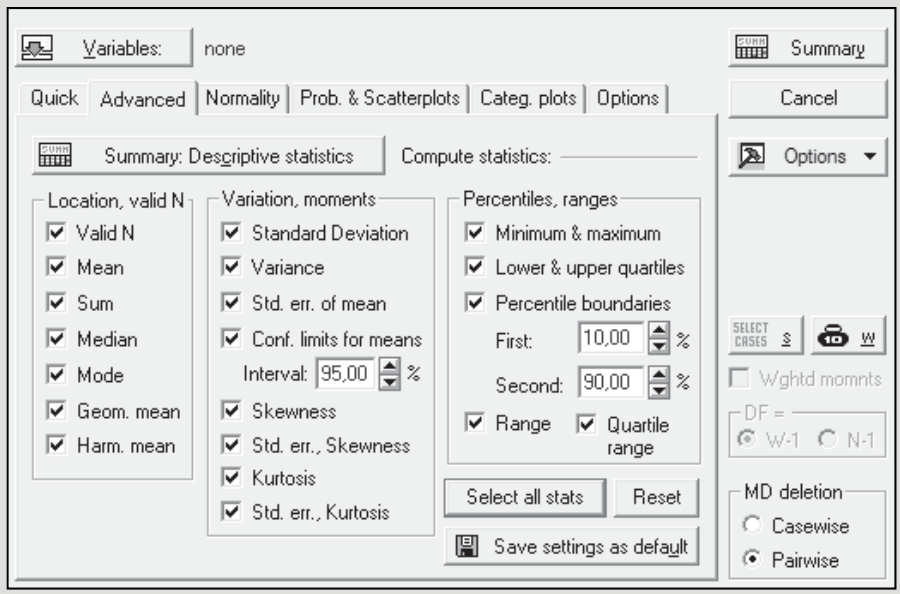


Paso 4. En la siguiente ventana seleccionamos «Descriptive statistics».

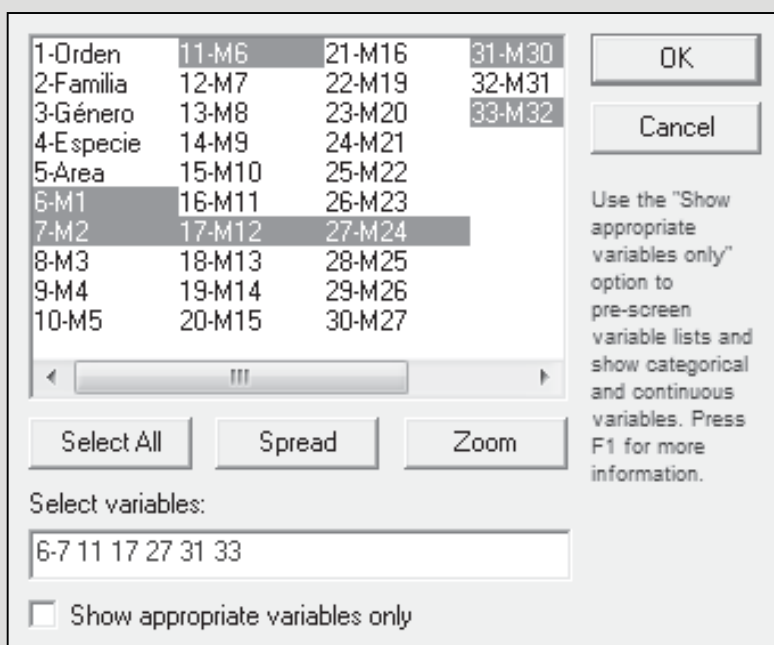


CUADRO II.7. *(Continuación)*[STATISTICA]

Paso 5. A continuación en la ventana que aparece se selecciona la pestaña «Advanced». Se puede marcar unos pocos estadísticos o pulsar «Select all stats» para seleccionarlos todos.



Paso 6. En la ventana anterior pulsamos en «Variables» y en la ventana que aparece seleccionamos las variables que deseamos analizar.

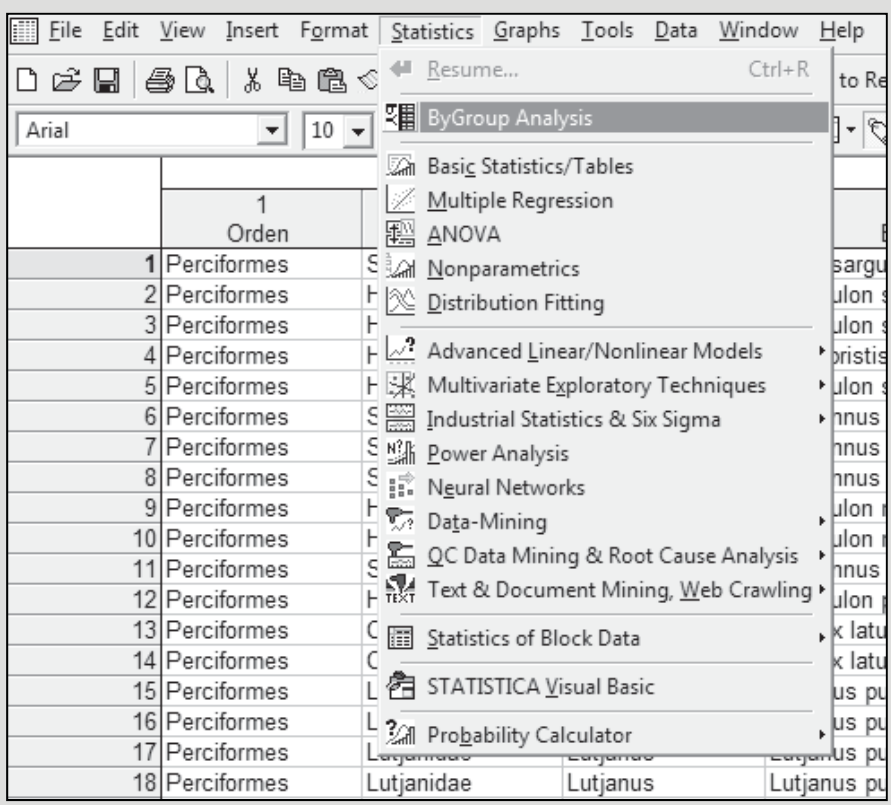


CUADRO II.7. *(Continuación)*[STATISTICA]

Paso 7. Por último, en la ventana del paso 5 pulsamos «Summary» y obtenemos los resultados.

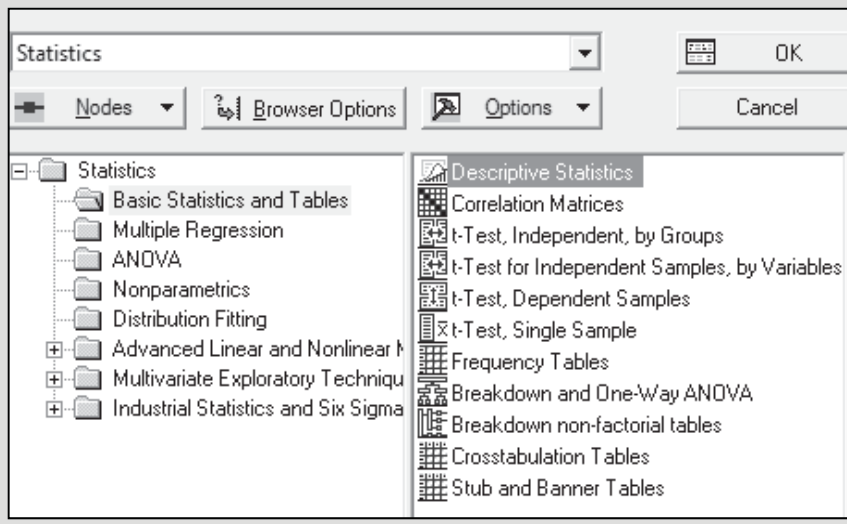
Variable	Descriptive Statistics (Cuadro II.7.sta) Include condition: v4="Lutjanus purpureus"				
	Valid N	Mean	Confidence -95,000%	Confidence +95,000%	Geometric Mean
M1	5	302,9860	291,4508	314,5212	302,8707
M2	5	35,6780	32,3144	39,0416	35,5971
M6	5	117,8860	113,0337	122,7383	117,8344
M12	5	152,0020	142,1216	161,8824	151,8362
M24	5	39,7860	38,0710	41,5010	39,7667
M30	5	22,9140	21,4421	24,3859	22,8903
M32	5	17,3180	14,9752	19,6608	17,2393

Paso 8. STATISTICA permite hacer la estadística descriptiva de determinado grupo de datos de otra forma. Se selecciona «Statistics» ⇨ «ByGroup Analysis».

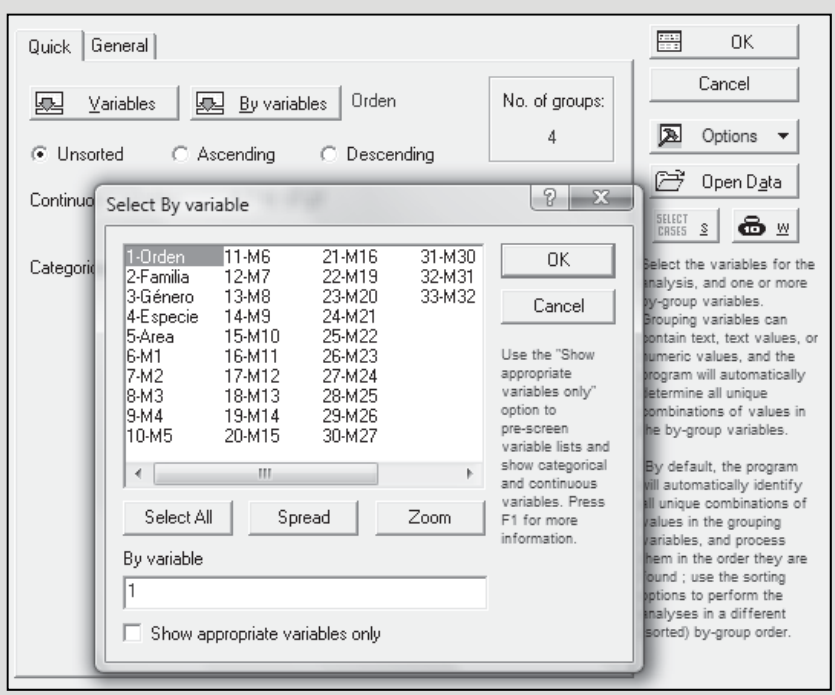


CUADRO II.7. *(Continuación)*[STATISTICA]

Paso 9. En la ventana que aparece se selecciona «Basic Statistics and Tables» ⇨ «Descriptive Statistics».

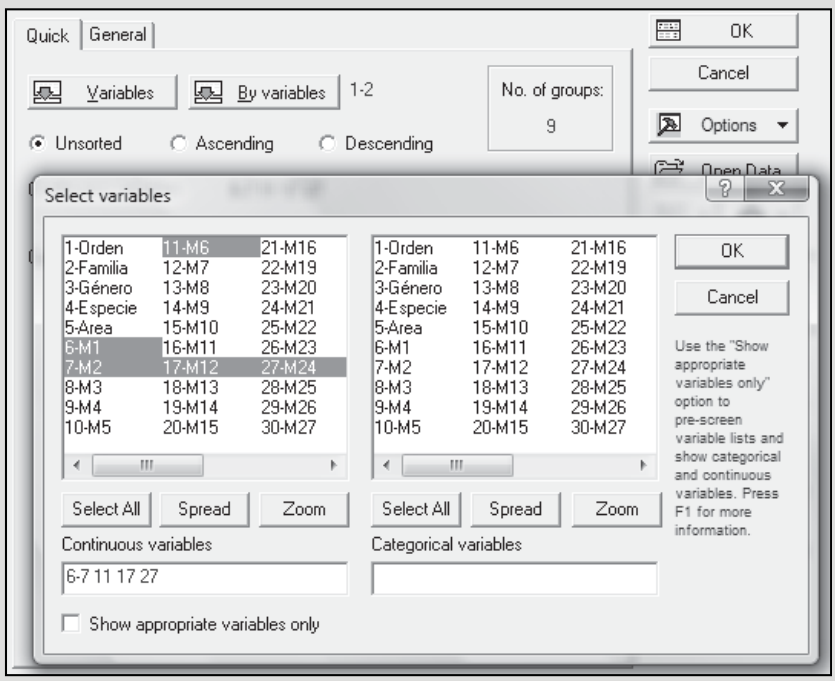


Paso 10. En la nueva ventana, en la pestaña «Quick» pulsamos el botón «By variables» y seleccionamos las variables de agrupación.

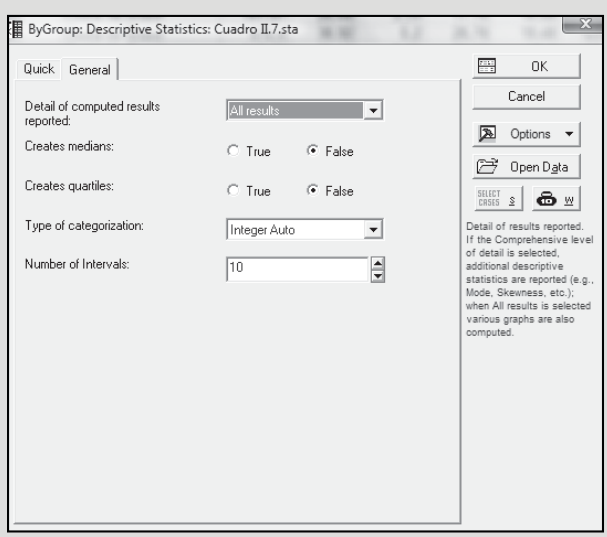


CUADRO II.7. *(Continuación)* *STATISTICA*

Paso 11. En el botón «Variables» seleccionamos las variables que se quieren analizar.

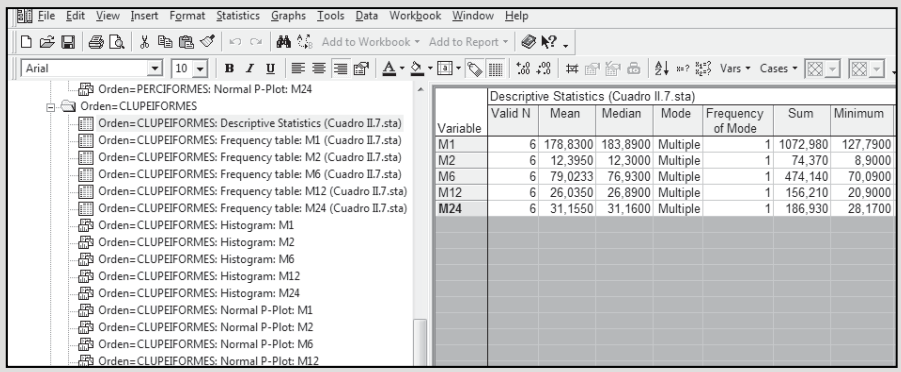


Paso 12. En la pestaña «General» seleccionamos «All results» para poder ver el máximo posible de parámetros estadísticos y pulsamos «OK».



CUADRO II.7. *(Continuación)* *STATISTICA*

Paso 13. Para poder ver los resultados es necesario pulsar en la ventana de la izquierda en cada una de las combinaciones de las agrupaciones.



Descriptive Statistics (Cuadro II.7.sta)

Variable	Valid N	Mean	Median	Mode	Frequency of Mode	Sum	Minimum
M1	6	178,8300	183,8900	Multiple	1	1072,980	127,7900
M2	6	12,3950	12,3000	Multiple	1	74,370	8,9000
M6	6	79,0233	76,9300	Multiple	1	474,140	70,0900
M12	6	26,0350	26,8900	Multiple	1	156,210	20,9000
M24	6	31,1550	31,1600	Multiple	1	186,930	28,1700

Capítulo III

DISTRIBUCIÓN

Otra información importante es la relacionada con la distribución que tiene una variable. Se puede dar el caso de que dos variables tengan exactamente la misma media y la misma dispersión, pero que el tipo de distribución sea completamente diferente. Por lo tanto, además de la información sobre las medidas de posición y dispersión, que se explicaron en el Capítulo II, es necesario también conocer la distribución de los valores de la variable.

Para estudiar la distribución de una variable, e incluso comparar medias y varianzas entre variables, lo que se hace es comparar las frecuencias de los valores de la variable con las probabilidades resultantes de modelos teóricos de distribuciones. El modelo teórico de distribución a utilizar variará en función del tipo de variable con la que estemos trabajando.

III.1. Tipos de variables

Existen básicamente dos tipos de variables: *cualitativas* y *cuantitativas*. Las primeras no se expresan numéricamente (sexo, especie a la que pertenece un individuo, provincia de nacimiento) aunque pueden utilizarse códigos numéricos para representar sus valores (por ejemplo en la variable sexo, en vez de macho y hembra se pueden nombrar como 1 y 2). Las cuantitativas se expresan directamente en términos numéricos (número de hojas de una planta, edad, longitud, temperatura, etc.).

Las variables cualitativas a su vez pueden ser de dos tipos: *nominales*, si sus valores no están ordenados de modo natural (lugar de nacimiento, especie), y *ordinales* si sus valores tienen un orden (por ejemplo una variable "toxicidad" que toma los valores nada, poco, bastante y muy tóxico).

Las variables cuantitativas pueden ser también de dos tipos: *discretas* y *continuas*. Son discretas cuando solo pueden tomar valores específicos, y no es posible ningún valor entre dos consecutivos (número de hijos que puede tener una familia, número de hojas de una planta, etc.). En el caso de variables discretas a cada valor de la variable le corresponde una probabilidad, de la cual dependerá el número de veces que se repita ese valor en relación con los valores restantes. A la correspondencia entre los valores y sus respectivas probabilidades se le llama *función de probabilidad*.

Las variables cuantitativas continuas son aquellas que pueden tomar cualquier valor a lo largo de un continuo, de modo que no hay valores consecutivos, ya que entre dos valores cualesquiera siguen existiendo infinitos valores posibles (temperatura, longitud). Las variables continuas se pueden agrupar en categorías, pero de una forma arbitraria. Por ejemplo, la variable altura se puede dividir en categorías como pequeño, normal y alto, y los límites de cada una de estas categorías se pueden establecer de forma arbitraria. A diferencia de lo que ocurría con las variables discretas, para las variables continuas no es útil establecer la correspondencia entre valores y probabilidades. Lo que se hace es calcular la probabilidad contenida en un determinado segmento o rango de valores,

que dividida por la amplitud del segmento es la densidad media de probabilidad del segmento, a partir de la cual se determina la densidad de probabilidad para cada valor. La correspondencia entre los valores y su respectiva densidad de probabilidad se llama *función de densidad*.

La distinción entre estos cuatro tipos de variables es importante por varios motivos:

1. El cálculo de algunas medidas de posición o de dispersión no tiene sentido con variables cualitativas, por ejemplo en el caso de la variable sexo.
2. Para la aplicación correcta de técnicas de análisis estadísticos: así la mayoría de las pruebas no paramétricas requieren que la variable sea al menos ordinal, y muchos métodos de análisis multivariante exigen que las variables sean cuantitativas (por ejemplo análisis factorial o análisis discriminante).

En este capítulo vamos a estudiar las funciones de probabilidad y de densidad de las distribuciones para variables discretas y continuas.

III.2. Distribuciones para variables continuas

III.2.1. NORMAL

La distribución Normal es una de las más importantes porque se observa en muchas series de datos y, además, distintos tipos de análisis estadísticos presentan como condición para poder aplicarse a la serie de datos, que ésta presente este tipo de distribución. La forma de la distribución Normal varía en función de la media (μ) y la desviación típica (σ) como se observa en la Figura III.1.

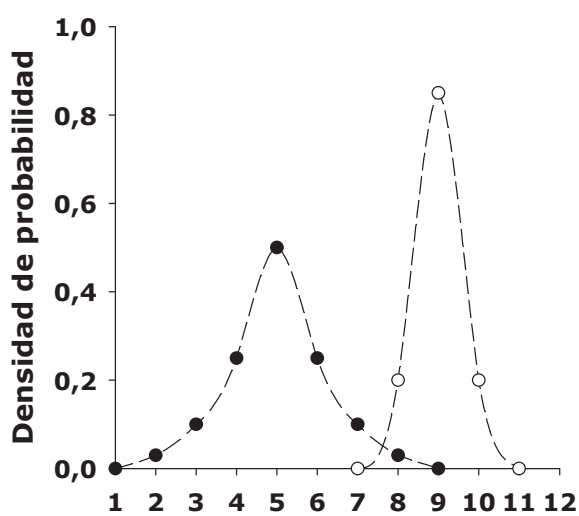


Figura III.1. Distribuciones normales con diferentes medias y desviaciones típicas.

Varios matemáticos intervinieron en su desarrollo, entre los que figura el matemático y astrónomo Carl Friedrich Gauss (1777–1855), por lo cual a veces es llamada en su honor la campana o distribución de Gauss.

La función de densidad de la distribución Normal se describe por medio de la siguiente fórmula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2}$$

donde *f(x)* es la densidad de probabilidad del valor *x*, μ es la media y σ es la desviación típica.

III.2.1.1. Aplicaciones de la distribución Normal

La tabla de la distribución Normal N(0,1), es decir con $\mu = 0$ y $\sigma = 1$ (Tabla 1 del Apéndice I y archivo **Tabla 1.doc** en la carpeta Apéndice), nos permite calcular probabilidades relativas a cualquier otra distribución Normal con diferente μ y σ**.** Para ello, basta tipificar la variable, es decir, calcular el valor *Z* (unidades de desviación con respecto a la media) correspondiente a los valores *x* indicados mediante la operación:

$$Z = \left(\frac{x - \mu}{\sigma}\right)$$

Este valor *Z* que se calcula a partir de la variable *X* nos permite obtener en las tablas las probabilidades correspondientes a cualquier intervalo (véase Cuadro III.1).

CUADRO III.1. Usos de la distribución Normal

EJEMPLO. La longitud del cuerpo de una especie de una determinada población se distribuye según una Normal de $\mu = 10{,}8$ cm y $\sigma = 3{,}7$ cm.

1. Calcular la probabilidad de que un individuo tenga un tamaño mayor de 8,9 cm.

Paso 1.1. Cálculo del valor *Z*

$$Z = \left(\frac{x - \mu}{\sigma}\right) = \frac{8{,}9 - 10{,}8}{3{,}7} = -0{,}51$$

Paso 1.2. El paso siguiente es consultar la tabla de la distribución Normal N(0,1) (Tabla 1 del Apéndice I o archivo **Tabla 1.doc**), que para un *Z* de 0,51 (se busca siempre el valor absoluto), el valor en la tabla es de 0,695.

CUADRO III.1. *(Continuación)*

Paso 1.3. Se pueden dar 4 situaciones:

1. El valor de x es menor que μ (en nuestro ejemplo 10,8 cm) y se pide la probabilidad de que sea superior a x. En este caso la probabilidad será el valor obtenido en la **Tabla 1.doc**.
2. El valor de x es menor que μ y se pide la probabilidad de que sea inferior a x. En este caso la probabilidad será 1 menos el valor obtenido en la **Tabla 1.doc**.
3. El valor de x es mayor que μ y se pide la probabilidad de que sea inferior a x. En este caso la probabilidad será el valor obtenido en la **Tabla 1.doc**.
4. El valor de x es mayor que μ y se pide la probabilidad de que sea superior a x. En este caso la probabilidad será 1 menos el valor obtenido en la **Tabla 1.doc**.

En nuestro ejemplo, x (8,9) es menor que μ (10,8) y se pide la probabilidad de que sea superior a x. Por lo tanto, sería la situación 1 y la probabilidad de elegir un individuo al azar que tenga más de 8,9 cm de longitud será de 0,695.

La probabilidad de que un individuo fuese menor de 8,9 cm sería:

$$1 - 0{,}695 = 0{,}305$$

2. Calcular la probabilidad de que un individuo tenga un tamaño entre 8,9 y 12,4 cm.

$$Z = \left(\frac{x - \mu}{\sigma}\right) = \frac{8{,}9 - 10{,}8}{3{,}7} = -0{,}51$$

La probabilidad de que sea menor de 8,9 cm es 0,305.

$$Z = \frac{x - \mu}{\sigma} = \frac{12{,}4 - 10{,}8}{3{,}7} = 0{,}43$$

La probabilidad de que sea menor de 12,4 cm es 0,6664.

Por lo tanto, la probabilidad de que un individuo tenga un tamaño entre 8,9 y 12,4 cm es **0,6664 – 0,305** = 0,3614.

3. Calcular la probabilidad de que un individuo tenga un tamaño entre 11,9 y 12,4 cm.

La probabilidad de que sea menor de 11,9 cm es 0,6141.

La probabilidad de que sea menor de 12,4 cm es 0,6664.

CUADRO III.1. *(Continuación)*[R]

Por lo tanto, la probabilidad de que un individuo tenga un tamaño entre 11,9 y 12,4 cm es 0,6664 – 0,6141 = 0,0523.

Aplicación con R.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro III.1.R.** Este archivo está en el CD que acompaña al libro en la carpeta Capítulo III. La siguiente ventana muestra las instrucciones del archivo **Cuadro III.1.R**.

```
# Cuadro III.1
# Usos de la distribución Normal
#########################################################
# Sección modificable por el usuario
#########################################################

# Valor que se quiere calcular la probabilidad en la normal
valor<-8.9
# Media de la distribución normal
media<-10.8
# Desviación estándar de la distribución normal
desviacion.estandar<-3.7
# ¿Se quiere obtener la probabilidad mayor?
prob.mayor<-"No"
```

Paso 1. Lo único que hay que hacer es cambiar la sección modificable por el usuario. En concreto, especificar el valor para el cual se quiere estimar la probabilidad, la media, la desviación estándar y si se quiere obtener la probabilidad mayor («Si») o menor («No») que el valor.

```
# Media de la distribución normal
media<-10.8
# Desviación estándar de la distribución normal
desviacion.estandar<-3.7
# ¿Se quiere obtener la probabilidad mayor?
prob.mayor<-"No"
```

Paso 2. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*.

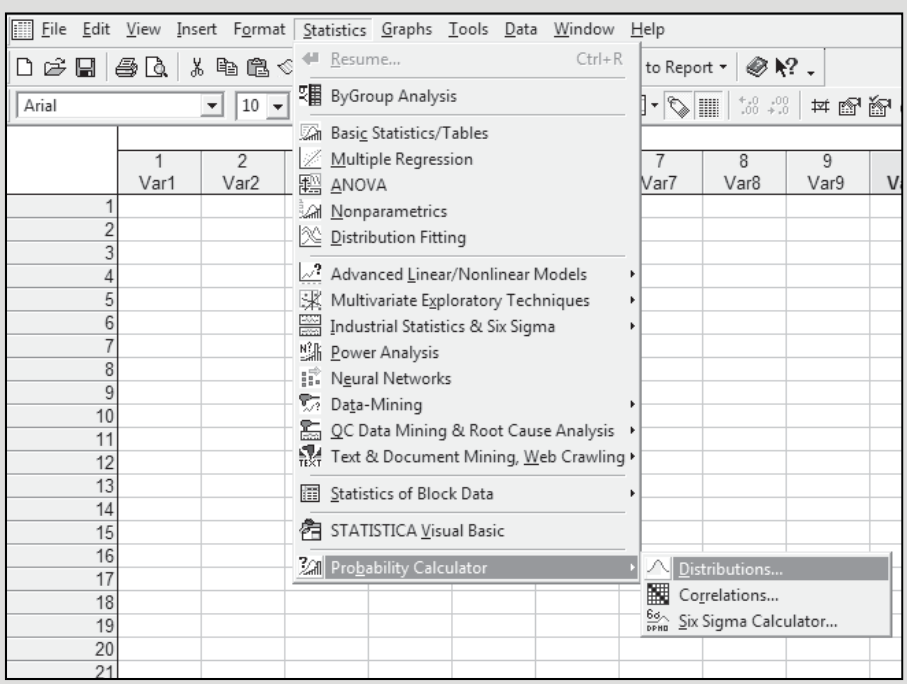
CUADRO III.1. *(Continuación)*[STATISTICA]

Paso 3. En la siguiente ventana se muestra el resultado. El cálculo con R puede diferir un poco con el cálculo utilizando la tabla de distribución Normal, debido al redondeo que se debe hacer al usar la tabla.

```
> ##########################################
> # Sección que muestra los resultados
> ##########################################
> print("Valor de la probabilidad normal")
[1] "Valor de la probabilidad normal"
> probabilidad
[1] 0.3037961
```

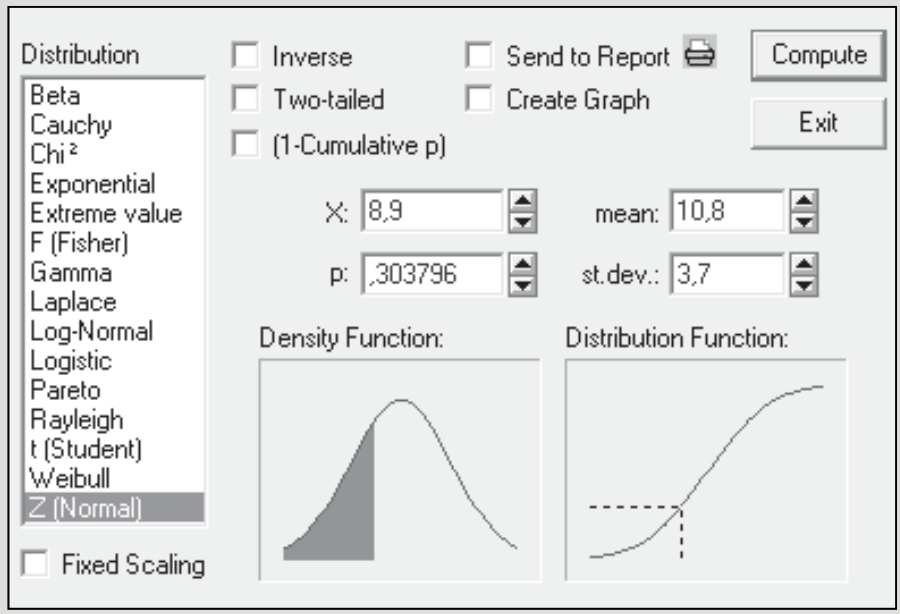
Aplicación con STATISTICA.

Paso 1. El programa STATISTICA permite realizar los cálculos que se han mostrado anteriormente. Para ello hay que entrar en «Statistics» ⇨ «Probability Calculator» ⇨ «Distributions».



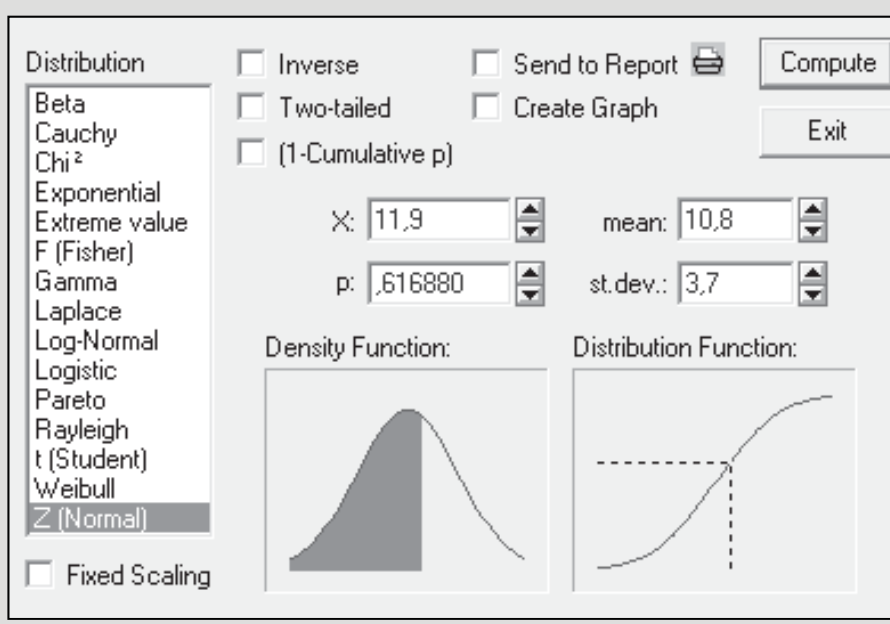
Paso 2. Nos aparecerá la siguiente ventana donde seleccionamos la distribución Normal «Z (Normal)».

CUADRO III.1. *(Continuación)*[STATISTICA]



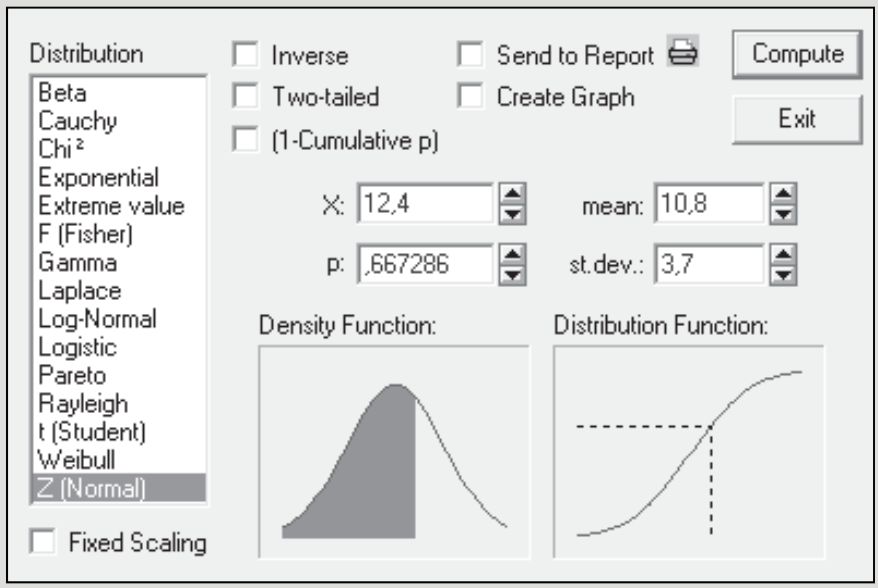
Paso 3. En el ventana anterior hay que introducir la media «mean», la desviación estándar «st.dev.» y el valor de «*X*». Después de introducir todos los datos se pulsa «Compute». Como ejemplo se muestra el caso del problema 1.

En el caso del problema 3, primero se calcula la probabilidad de que sea menor de 11,9 cm, como se muestra en la siguiente ventana.



CUADRO III.1. *(Continuación)*[SPSS]

Luego se calcula la probabilidad de que sea menor de 12,4 cm.

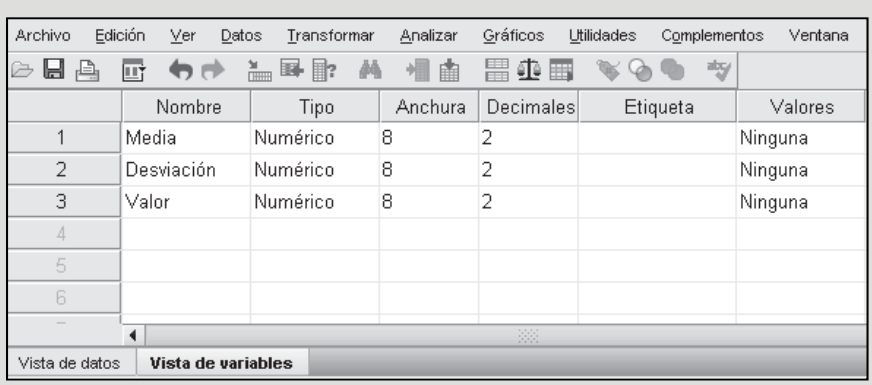


El resultado, como se mostró anteriormente, sería la diferencia entre ambas probabilidades:

$$0{,}667286 - 0{,}61688 = 0{,}0504$$

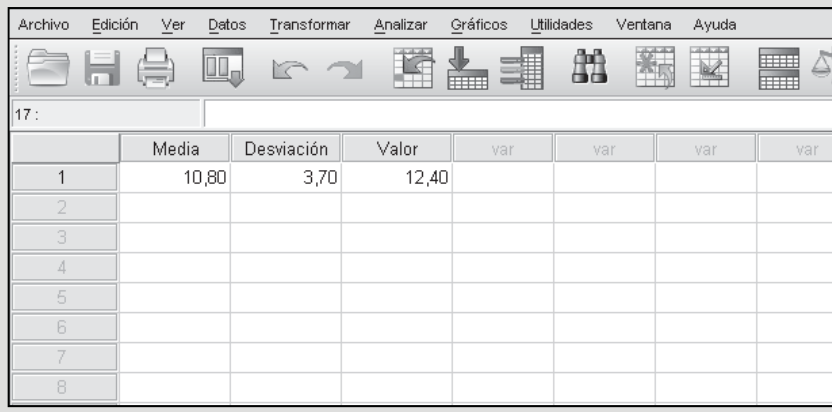
Aplicación con SPSS.

Paso 1. Una vez dentro de SPSS, en la pestaña «Vista de variables», en el apartado «Nombre» se crean tres variables: la media, la desviación y el valor para el cual vamos a calcular la probabilidad. En «Decimales» es posible especificar el número de decimales que se desea para cada variable.

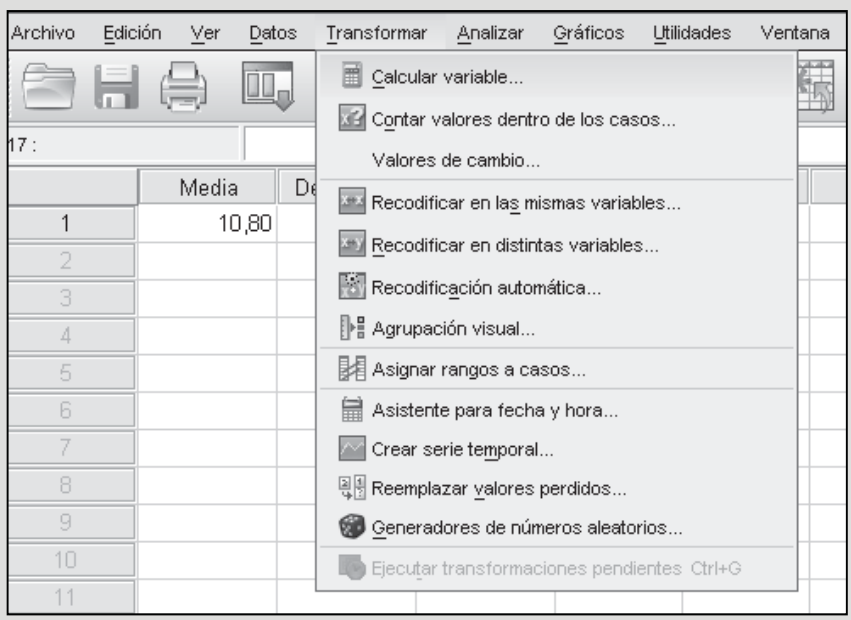

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Complementos Ventana
Nombre Tipo Anchura Decimales Etiqueta Valores
1 Media Numérico 8 2 Ninguna
2 Desviación Numérico 8 2 Ninguna
3 Valor Numérico 8 2 Ninguna
4
5
6
Vista de datos
Vista de variables

CUADRO III.1. *(Continuación)*[SPSS]

Paso 2. En la pestaña «Vista de datos», aparece la matriz donde se introducen los valores de cada variable.



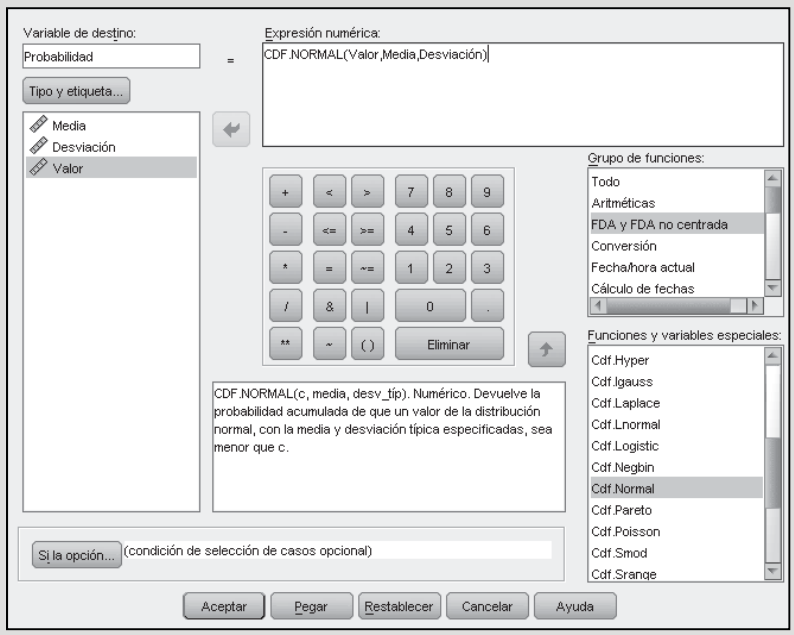
Paso 3. A continuación se entra en «Transformar» ⇨ «Calcular variable…».



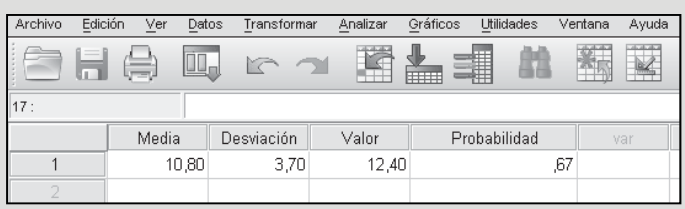
Paso 4. En la ventana que aparece, se introduce en «Variable de destino:» el nombre de la variable donde se incluirá el cálculo de la probabilidad, en este caso "Probabilidad". Seleccionando «FDA y FDA no centrada» aparece un listado donde se selecciona «Cdf.Normal». Pulsando sobre la flecha que está al lado del botón «Eliminar», se pasa esta función al cuadro de «Expresión numérica».

CUADRO III.1. *(Continuación)*[SPSS]

En la función aparecen unos interrogantes que hay que sustituir por el valor, media y desviación, que se seleccionan en el cuadro de variables a la izquierda.



Paso 5. Al pulsar «Aceptar» aparece el resultado de la probabilidad.



III.2.1.2. Asimetría

Las medidas de asimetría tienen como finalidad el determinar si una variable se distribuye de forma simétrica con respecto a un valor central, o bien si la distribución de los datos tiene una forma diferente del lado derecho que del lado izquierdo.

La distribución es simétrica si al lado derecho y al lado izquierdo del valor central la distribución de los datos es la misma. La distribución tiene asimetría positiva si las frecuencias más altas se encuentran en el lado izquierdo de la media, mientras que en el derecho hay frecuencias más pequeñas. La asimetría es negativa cuando las frecuencias más pequeñas están en el lado izquierdo (Figura III.2).

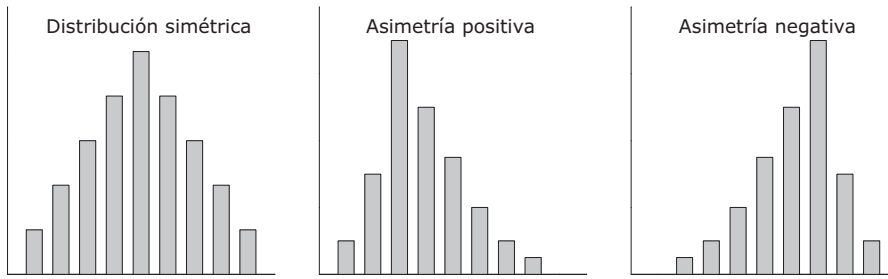


Figura III.2. Diferentes tipos de asimetrías.

Como valor central de referencia se suele usar la media aritmética, aunque también se puede utilizar la mediana.

Existen muchas formas de medir la simetría y una de ellas es la fórmula que se muestra a continuación (coeficiente de asimetría de Fisher), la cual utiliza como valor central la media aritmética (Cuadro III.2 y archivo **Cuadro III.2.xls**):

$$Asimetría = \frac{n\left(\sum_{i=1}^{n}(x_i - \bar{x})^3\right)}{(n-1)(n-2)\sigma^3}$$

siendo σ la desviación típica estimada o cuasidesviación típica muestral. Cuando el valor es cercano a cero la distribución es simétrica, la asimetría es positiva cuando el valor es mayor que cero y la asimetría es negativa cuando el valor es menor que cero.

III.2.1.3. Apuntamiento o curtosis

Las medidas de curtosis estudian la distribución de los datos en la zona central de la serie. La mayor o menor concentración de frecuencias alrededor de la media y en la zona central de la distribución dará lugar a una distribución más o menos apuntada. Por esta razón, a las medidas de curtosis se les llama también de apuntamiento o concentración central. Las medidas de curtosis se aplican a distribuciones cercanas a la normalidad, es decir, unimodales simétricas o con ligera asimetría.

Al igual que ocurría con la asimetría, existen diversas formas de estimar la curtosis de una distribución de datos, pero una de las más usadas es el coeficiente de apuntamiento de Fisher que se muestra a continuación (Cuadro III.2 y archivo **Cuadro III.2.xls**).

$$Curtosis = \frac{n(n+1)\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)(n-1)}{(n-1)(n-2)(n-3)\sigma^4}$$

Cuando el valor es positivo se dice que la distribución está más apuntada que la Normal y se denomina leptocúrtica (Figura III.3).

Cuando el valor es cero o cercano a cero, la distribución tiene el mismo apuntamiento que la distribución Normal y se le denomina mesocúrtica (Figura III.3).

Por último, cuando el valor es negativo, el apuntamiento es menor que el de la distribución Normal y se le denomina platicúrtica (Figura III.3).

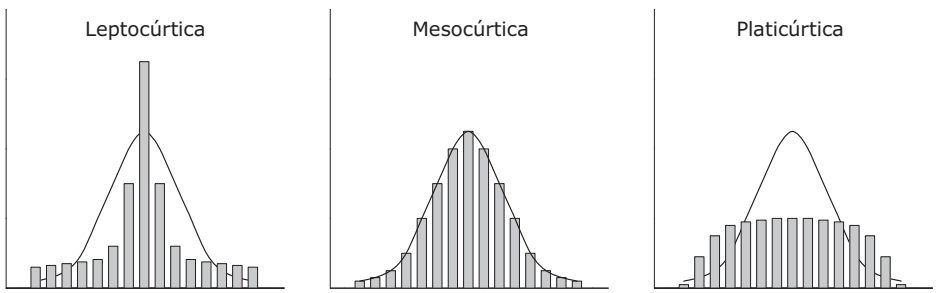


Figura III.3. Tipos de curtosis (barras) en comparación con la distribución Normal (línea continua).

CUADRO III.2. Asimetría y curtosis

Vamos a utilizar la misma serie de datos que en el Cuadro II.1 del Capítulo II. Estos cálculos se pueden realizar utilizando R, STATISTICA y SPSS como se mostró en el Capítulo II.

1. Asimetría

$$Asimetría = \frac{20 * ((22 - 29{,}3)^3 + ...(39 - 29{,}3)^3)}{19 * 18 * 5{,}39^3} = 0{,}197$$

2. Curtosis

$$Curtosis = \frac{20 * (21) * 24677{,}47 - 3 * 552{,}2 * 552{,}2 * 19}{19 * 18 * 17 * 5{,}39^4} = -1{,}43$$

III.2.2. *t* DE STUDENT

Cuando una variable sigue una distribución Normal, la media de una muestra aleatoria de esa variable también tiene distribución Normal, y su media es la media poblacional desconocida μ. Esto puede ser utilizado para estimar μ. Sin embargo, a menudo no se conoce la desviación típica de la población σ (solo se trabaja con una muestra de individuos del total de la población) y, además, puede ocurrir que el número de observaciones de la muestra sea pequeño (menor de 30).

En estos casos, se puede utilizar la cuasidesviación típica de la muestra (*s*) junto con la distribución *t* de Student:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

La función de densidad de probabilidad de la distribución *t* de Student viene dada por la siguiente expresión:

$$f(x) = \frac{1}{\sqrt{\upsilon\pi}} \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\Gamma\frac{\upsilon}{2}} \left(1 + \frac{x^2}{\upsilon}\right)^{-\left(\frac{\upsilon+1}{2}\right)}$$

La distribución *t* de Student puede tener diferentes formas dependiendo de los grados de libertad υ (Figura III.4). La apariencia general de la distribución *t* es similar a la de la distribución Normal estándar. Sin embargo, la distribución *t* tiene colas más amplias que la Normal, es decir, la probabilidad de las colas es mayor que en la distribución Normal. La distribución *t* se transforma en una distribución Normal cuando el número de datos tiende a infinito. Los valores críticos de distintos niveles de significación y distintos grados de libertad se muestran en la Tabla 2 del Apéndice I o archivo **Tabla 2.doc**.

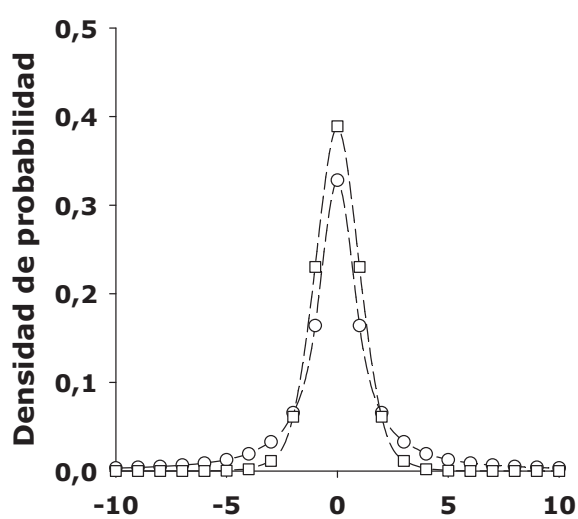


Figura III.4. Funciones de densidad de la distribución de Student para 1 (○) y 10 (□) grados de libertad.

Las aplicaciones de la distribución *t* de Student en la inferencia estadística son:

1. Para estimar, mediante intervalos de confianza, la media poblacional (Capítulo IV).
2. Estimar y probar hipótesis sobre una diferencia de medias (Capítulo IX).

Las hipótesis o supuestos para poder aplicar la *t* de Student son que en cada grupo la variable estudiada siga una distribución Normal y que la dispersión en ambos grupos sea homogénea (hipótesis de homocedasticidad = igualdad de varianzas) aunque, como veremos en el Capítulo IX, este estadístico también se puede usar sin asumir igualdad de varianzas.

III.2.3. CHI-CUADRADO

La función de densidad de la distribución Chi-cuadrado (χ^2) se describe por medio de la siguiente expresión:

$$f(x) = \frac{1}{2^{\frac{\upsilon}{2}}\Gamma\left(\frac{\upsilon}{2}\right)} e^{-\frac{X}{2}} X^{\left(\frac{\upsilon}{2}-1\right)}$$

donde υ son los grados de libertad y *x* no es negativo.

A diferencia de lo que ocurría con la distribución Normal, debido a que la distribución χ^2 depende de los grados de libertad, no existe una curva típica sino que la distribución χ^2 puede tener diferentes formas dependiendo de los grados de libertad (Figura III.5).

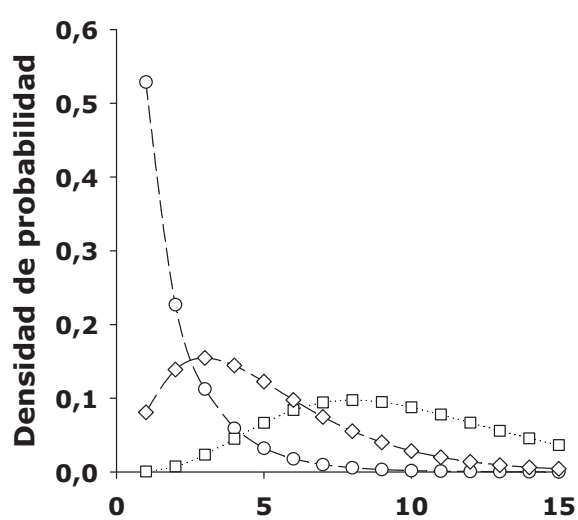


Figura III.5. Curvas de densidad de probabilidad de la distribución χ^2 para 1 (○), 5 (◇) y 10 (□) grados de libertad.

El valor de la variable que deja a su derecha un área α bajo la curva de densidad se llama el punto crítico correspondiente al nivel de significación α. Están tabulados los puntos críticos de distintos niveles de significación y distintos grados de libertad (Tabla 3 del Apéndice I o archivo **Tabla 3.doc**).

Son tres las aplicaciones principales que tiene la distribución χ^2: test de bondad de ajuste, test de independencia y test de homogeneidad. Además, esta distribución juega un papel importante en otras muchas pruebas estadísticas.

1. El test de bondad de ajuste consiste en el planteamiento de hasta qué punto una muestra se puede considerar como perteneciente a una población con una distribución teórica ya conocida. Es un método que se utiliza frecuentemente para determinar si una serie de datos presenta una distribución Normal, de Poisson, etc. (véase capítulo VI).
2. El test de independencia determina si dos caracteres *X* e *Y* de una población son dependientes o independientes. Por ejemplo, queremos determinar si la supervivencia de los descendientes de las hembras de una población es independiente o dependiente de la cantidad diaria que reciben de alimento las hembras (véase Capítulo VIII para variables cualitativas y Capítulo X para variables cuantitativas).
3. El test de homogeneidad permite determinar si varias muestras que estudian el mismo carácter *A* han sido tomadas o no de la misma población, respecto de dicha característica *A*. Por ejemplo, hemos seleccionado varios grupos de individuos de una población a los que hemos sometido a la misma dosis de diferentes metales y queremos determinar si los metales afectan de forma diferente a la supervivencia de los individuos (véase Capítulo VII para variables cualitativas y Capítulo IX para variables cuantitativas).

III.2.4. *F* DE FISHER-SNEDECOR

La función de densidad de probabilidad de la distribución *F* de Fisher-Snedecor viene dada por la siguiente expresión:

$$f(x) = \frac{\Gamma\left(\frac{\upsilon+\omega}{2}\right)}{\Gamma\frac{\upsilon}{2}\Gamma\frac{\omega}{2}} \upsilon^{\frac{\upsilon}{2}} \omega^{\frac{\omega}{2}} \frac{x^{\frac{\upsilon}{2}-1}}{(\omega+\upsilon x)^{\frac{\upsilon+\omega}{2}}}$$

donde υ y ω son los grados de libertad del numerador y denominador respectivamente, siendo *x* no negativo. Al depender de dos tipos de grados de libertad, la función de densidad puede tener diversas formas (Figura III.6).

Los valores críticos de la distribución *F* Fisher-Snedecor con distintos niveles de significación y distintos grados de libertad se muestran en la Tabla 4 del Apéndice I o en el archivo **Tabla 4.doc**.

Esta distribución se usa principalmente en dos tipos de situaciones, requiriéndosele en ambos casos que la distribución de las variables sea Normal:

1. Para probar si dos muestras provienen de poblaciones que poseen varianzas iguales. Esta prueba es útil para determinar si una población Normal tiene una mayor variación que la otra y es importante ya que,

como veremos en el capítulo IX, a la hora de comparar medias, varios estadísticos presentan como requisito la homogeneidad de varianzas.

2. También se aplica cuando se trata de comparar simultáneamente varias medias poblacionales (véase Capítulo VII para variables cualitativas y Capítulo IX para variables cuantitativas).

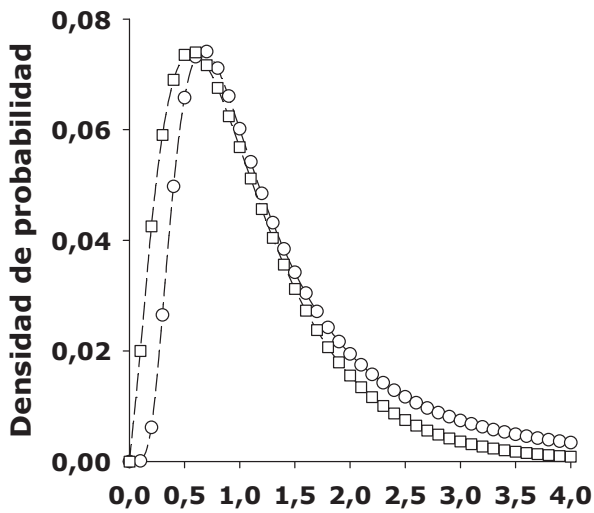


Figura III.6. Funciones de densidad de la distribución *F* Fisher-Snedecor para distintos grados de libertad $F_{(30,5)}$ (○) y $F_{(5,30)}$ (□).

III.3. Distribuciones para variables discretas

III.3.1. BINOMIAL

Una variable presenta una distribución binomial cuando sólo tiene dos posibles resultados: «éxito» y «fracaso», siendo la probabilidad de cada uno de ellos constante en una serie de repeticiones, es decir, ni la probabilidad de éxito ni la de fracaso cambian de una prueba a otra, y además el resultado de cada prueba es independiente del resultado de las demás pruebas. La probabilidad de éxito está representada por p y la probabilidad de fracaso se representa por $q = 1 - p$.

Como se comentó al inicio de este capítulo, en el caso de las variables discretas en lugar de la función de densidad se utiliza la función de probabilidad o de cuantía, que da una probabilidad para cada valor de la variable. La función de probabilidad binomial viene expresada por la siguiente fórmula:

$$f(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

En el Cuadro III.3 se muestran aplicaciones de la distribución binomial.

CUADRO III.3. Aplicación de la distribución binomial

EJEMPLO. La aplicación de un determinado tratamiento a un grupo de individuos que presentan un parásito produce una mejoría en un 67% de los casos. Si se aplica el tratamiento a 8 individuos:

El valor de p = 0,67 y, por tanto, el valor de q es 0,33.

¿Cuál es la probabilidad de que mejoren 7 individuos?

$$\frac{8!}{7!(8-7)!}0{,}67^{7}0{,}33^{8-7} = 0{,}16$$

¿Cuál es la probabilidad de que al menos mejoren 3 individuos?

La probabilidad de que al menos mejoren 3 individuos será 1 menos la probabilidad que mejoren 2 individuos menos la probabilidad de que mejore 1 individuo menos la probabilidad de que no mejore ninguno:

$$1-\frac{8!}{2!(8-2)!}0{,}67^{2}0{,}33^{8-2}-\frac{8!}{1!(8-1)!}0{,}67^{1}0{,}33^{8-1}-\frac{8!}{0!(8-0)!}0{,}67^{0}0{,}33^{8-0}$$

Probabilidad = 0,981

Aplicación con R. Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro III.3.R**, el cual se muestra en la siguiente ventana.

```
# Cuadro III.3
# Distribución binomial
#########################################################
# Sección modificable por el usuario
#########################################################

# Número total de individuos.
n=8
# Número de individuos
x=7
# Probabilidad de éxito.
p=0.67

# Opciones de cálculo de probabilidad
# 1. Exactamente la probabilidad P(X=x)
# 2. Probabilidad de que sea mayor igual p(X>=x)
# 3. Probabilidad de que sea estrictamente mayor p(X>x)
# 4. Probabilidad de que sea menor igual p(X<=x)
# 5. Probabilidad de que sea estrictamente menor p(X<x)
opcion<-1
```

CUADRO III.3. *(Continuación)*[R]

Paso 1. Hay que especificar, en la sección modificable por el usuario, el número total de individuos, el número de individuos para el cual se quiere calcular la probabilidad y la probabilidad de éxito.

Paso 2. Posteriormente, hay que indicar el tipo de probabilidad que se quiere calcular, seleccionando el número correspondiente (en este caso la primera opción).

```
# 1. Exactamente la probabilidad P(X=x)
# 2. Probabilidad de que sea mayor igual p(X>=x)
# 3. Probabilidad de que sea estrictamente mayor p(X>x)
# 4. Probabilidad de que sea menor igual p(X<=x)
# 5. Probabilidad de que sea estrictamente menor p(X<x)
opcion<-1
```

Paso 3. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*.

Paso 4. En la siguiente ventana se muestra el resultado.

```
> ########################################################
> # Sección que muestra los resultados
> ########################################################
> print("Valor de la probabilidad binomial")
[1] "Valor de la probabilidad binomial"
> probabilidad
[1] 0.1600028
```

III.3.2. HIPERGEOMÉTRICA

En la distribución hipergeométrica la variable también es aleatoria y dicotómica como en la distribución binomial, pero se diferencia de esta última en dos características importantes: la población es finita, mientras que en la binomial puede ser infinita y, además, la probabilidad *p* cambia -no es constante- ya que el resultado de cada prueba depende del resultado de las anteriores. La función de probabilidad viene expresada por la siguiente fórmula:

$$f(x) = \frac{\dfrac{N_p!}{x!(N_p - x)!}\dfrac{N_q!}{(n-x)!(N_q - n + x)!}}{\dfrac{N!}{n!(N-n)!}}$$

donde N_p y N_q son el número de elementos con probabilidad inicial p y q, respectivamente (es decir, $p = N_p/N$ y $q = N_q/N$), N el número total de elementos y n el número de elementos de la muestra extraída de los N de la población.

En el Cuadro III.4 se muestran aplicaciones de la distribución hipergeométrica.

CUADRO III.4. Aplicación de la distribución hipergeométrica

EJEMPLO. En un lote de 30 vacunas se ha comprobado que 8 están en mal estado. Ya se han utilizado 4 vacunas del lote.

¿Cuál es la probabilidad de que al menos una de las vacunas suministradas esté en malas condiciones?

Se calcula la probabilidad de que ninguna de las vacunas suministradas esté en mal estado

$$\frac{\frac{22!}{4!(22-4)!}\frac{8!}{(4-4)!(8-4+4)!}}{\frac{30!}{4!(30-4)!}} = 0{,}267$$

Por lo tanto, la respuesta sería 1 – 0,267= 0,733

¿Cuál es la probabilidad de que 3 de las vacunas suministradas estén en malas condiciones?

$$\frac{\frac{22!}{1!(22-1)!}\frac{8!}{(4-1)!(8-4+1)!}}{\frac{30!}{4!(30-4)!}} = 0{,}045$$

Aplicación con R.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro III.4.R.** Este archivo está en el CD que acompaña al libro en la carpeta Capítulo III. La siguiente ventana muestra las instrucciones del archivo **Cuadro III.4.R**.

```
# Cuadro III.4
# Distribución hipergeométrica.
#########################################################
# Sección modificable por el usuario
#########################################################

# Número total de elementos N
m=30
# Valor de Nq
n=8
# Valor del tamaño de muestra
k=4
# Valor para el que se quiere calcular la probabilidad
t=0

# Opciones de cálculo de probabilidad
# 1. Exactamente la probabilidad P(X=t)
# 2. Probabilidad de que sea mayor igual p(X>=t)
# 3. Probabilidad de que sea estrictamente mayor p(X>t)
# 4. Probabilidad de que sea menor igual p(X<=t)
# 5. Probabilidad de que sea estrictamente menor p(X<t)
opcion<-1
```

CUADRO III.4. *(Continuación)*[R]

Paso 1. Hay que especificar, en la sección modificable por el usuario, el número total de elementos *N*, el valor de N_q, el valor de *x* y el valor para el cual se quiere calcular la probabilidad.

```
# Número total de elementos N
m=30
# Valor de Nq
n=8
# Valor del tamaño de muestra
k=4
# Valor para el que se quiere calcular la probabilidad
t=0
```

Paso 2. Posteriormente, hay que indicar el tipo de probabilidad que se quiere calcular, seleccionando el número correspondiente (en este caso la primera opción).

```
# Opciones de cálculo de probabilidad
# 1. Exactamente la probabilidad P(X=t)
# 2. Probabilidad de que sea mayor igual p(X>=t)
# 3. Probabilidad de que sea estrictamente mayor p(X>t)
# 4. Probabilidad de que sea menor igual p(X<=t)
# 5. Probabilidad de que sea estrictamente menor p(X<t)
opcion<-1
```

Paso 3. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*.

Paso 4. En la siguiente ventana se muestra el resultado.

```
> ############################################################
> # Sección que muestra los resultados
> ############################################################
> print("Valor de la probabilidad hipergeométrica")
[1] "Valor de la probabilidad hipergeométrica"
> probabilidad
[1] 0.2669221
```

III.3.3. POISSON

Un proceso de Poisson es un proceso de sucesos independientes que se caracteriza por:

1. El número de sucesos en dos intervalos distintos siempre es independiente.
2. La probabilidad de que un suceso ocurra en un intervalo infinitesimal es proporcional a la longitud del intervalo.

3. La probabilidad de que ocurra más de un suceso en un intervalo muy pequeño *h* es 0.
4. Los sucesos son expresados por unidad de área, tiempo, etc.

La distribución de Poisson describe el número de sucesos en una unidad de tiempo de un proceso Poisson. Muchos fenómenos se modelan como un proceso de Poisson, por ejemplo el número de accidentes en una determinada zona de una carretera.

Las diferencias más importantes con respecto a la distribución binomial son que esta distribución se aplica a sucesos que pueden tener una probabilidad muy baja y, además, el tamaño de *n* es infinito. En algunos casos la distribución de Poisson se utiliza como aproximación a la Binomial cuando *n* es muy grande y, por tanto, es difícil el cálculo de la Binomial y, además, cuando la probabilidad de alguno de los eventos es muy baja. La función de probabilidad de la distribución de Poisson se expresa por la siguiente fórmula:

$$f(x) = \frac{\lambda^X}{x! e^{\lambda}}$$

donde λ es la media o promedio de sucesos por unidad de tiempo y x es la variable que indica el número de sucesos.

En el Cuadro III.5 se dan varios ejemplos de aplicación de la distribución de Poisson.

CUADRO III.5. Aplicación de la distribución de Poisson

EJEMPLO 1. La abundancia de una especie es de 23 individuos por 100 m^2. Como se trata de sucesos por unidad de área se utiliza una Poisson.

¿Cuál es la probabilidad de no encontrar ningún individuo en 25 m^2?

$$\lambda = \frac{23 individuos * 25 m^2}{100 m^2} = 5{,}75$$

La probabilidad que buscamos será: $\frac{5{,}75^0}{0! e^{5{,}75}} = 0{,}0031$

EJEMPLO 2. El número de tigres observados fue de 120 en 30 días.

Como se trata de sucesos por unidad de tiempo se utiliza una Poisson.

¿Cuál es la probabilidad de ver 5 tigres en 10 días?

$$\lambda = \frac{120 tigres * 10 días}{30 días} = 40$$

La probabilidad que buscamos será: $\frac{40^5}{5! e^{40}} = 3{,}62^{-12}$

CUADRO III.5. *(Continuación)*[R]

Aplicación con R.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro III.5.R.** Este archivo está en el CD que acompaña al libro en la carpeta Capítulo III. La siguiente ventana muestra las instrucciones del archivo **Cuadro III.5.R**.

```
# Cuadro III.5
# Distribución de Poisson.
#########################################################
# Sección modificable por el usuario
#########################################################

# Valor del parámetro de lambda
lambda <- 5.75
# Para el segundo ejemplo
# lambda <- 120*10/30

# Valor sobre el cual se quiere calcular la probabilidad
x<-0
# Valor sobre el cual se quiere calcular la probabilidad (segundo ejemplo)
# x<-5


# Opciones de cálculo de probabilidad
# 1. Exactamente la probabilidad P(X=x)
# 2. Probabilidad de que sea mayor igual p(X>=x)
# 3. Probabilidad de que sea estrictamente mayor p(X>x)
# 4. Probabilidad de que sea menor igual p(X<=x)
# 5. Probabilidad de que sea estrictamente menor p(X<x)
```

Paso 1. Hay que especificar el valor de λ y el valor para el cual se quiere calcular la probabilidad. El *script* está preparado para calcular el ejemplo 1, pero también se muestra para el caso del ejemplo 2. Solo es necesario quitar el símbolo #, para que se ejecute el segundo ejemplo.

```
# Valor del parámetro lambda
lambda <- 5.75
# Para el segundo ejemplo
# lambda <- 120*10/30

# Valor sobre el cual se quiere calcular la probabilidad.
x<-0
# Valor sobre el cual se quiere calcular la probabilidad (segundo ejemplo)
# x<-5
```

Paso 2. Posteriormente, hay que indicar el tipo de probabilidad que se quiere calcular, seleccionando el número correspondiente (en este caso la primera opción).

CUADRO III.5. *(Continuación)*[R]

```
# Opciones de cálculo de probabilidad
# 1. Exactamente la probabilidad P(X=x)
# 2. Probabilidad de que sea mayor igual p(X>=x)
# 3. Probabilidad de que sea estrictamente mayor p(X>x)
# 4. Probabilidad de que sea menor igual p(X<=x)
# 5. Probabilidad de que sea estrictamente menor p(X<x)
opcion<-1
```

Paso 3. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*.

Paso 4. En la siguiente ventana se muestra el resultado.

```
> ############################################################
> # Sección que muestra los resultados
> ############################################################
> print("Valor de la probabilidad de Poisson")
[1] "Valor de la probabilidad de Poisson"
> probabilidad
[1] 0.003182781
```

Capítulo IV

INTERVALOS DE CONFIANZA

En capítulos anteriores se ha descrito la forma de estimar determinados parámetros que nos permiten tener información sobre las variables que estamos estudiando, como por ejemplo los distintos tipos de media, varianza, desviación típica, etc. En la mayoría de los casos estos parámetros se estiman a partir de los datos de una muestra, y no del conjunto total de la población. Para medir la fiabilidad de la estimación es necesario calcular el *intervalo de confianza* del parámetro.

El intervalo de confianza de un parámetro al nivel de confianza 1 - α es aquel que cumple la propiedad de que la probabilidad de que sus extremos tomen valores tales que el parámetro esté comprendido entre ellos es igual a 1 - α (Viedma, 1989). Los extremos del intervalo de confianza de un parámetro se denominan límites de confianza.

IV.1. Intervalo de confianza de la media de una población Normal

IV.1.1. DESVIACIÓN TÍPICA CONOCIDA

En el supuesto de que se conozca la desviación típica de la población, el intervalo de confianza de la media al nivel de 1 - α ($I_{\mu}^{1-\alpha}$) se calcula por medio del intervalo estimador siguiente:

$$I_{\mu}^{1-\alpha} = \left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

Los valores críticos que se utilizan son las áreas bajo la curva Normal estándar (Tabla 1 del Apéndice I o archivo **Tabla 1.doc**).

En el Cuadro IV.1 se da un ejemplo de cómo estimar ese tipo de intervalo de confianza.

IV.1.2. DESVIACIÓN TÍPICA DESCONOCIDA

Lo más frecuente es que estemos trabajando con unos cuantos elementos del total que forma la población. Esto quiere decir que no conocemos la desviación típica del conjunto de toda la población. En este tipo de situación, el intervalo de confianza se calcula de dos formas diferentes dependiendo del tamaño de la muestra.

IV.1.2.1. Tamaño de muestra pequeño (< 30)

Si el tamaño de la muestra es pequeño, en vez de utilizar la distribución Normal y, por tanto, el estadístico *Z*, lo que se utiliza es la distribución *t* de Student con *n* - 1 grados de libertad. El intervalo de confianza de la media al nivel de confianza 1 - α ($I_{\mu}^{1-\alpha}$) es:

$$I_{\mu}^{1-\alpha} = \left(\bar{x} - t_{\frac{\alpha}{2};n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2};n-1} \frac{s}{\sqrt{n}} \right)$$

Se utilizan los valores críticos de la distribución *t* de Student, que se pueden ver en la Tabla 2 del Apéndice I o archivo **Tabla 2.doc**.

En el Cuadro IV.1 se da un ejemplo de cómo estimar ese tipo de intervalo de confianza.

IV.1.2.2. Tamaño de muestra grande (≥ 30)

Cuando el tamaño de la muestra tiende a infinito, la *t* de Student tiende a una distribución Normal estándar. Por ello, la mayoría de las tablas de la distribución *t* tienen solamente los valores correspondientes a un número de grados de libertad entre 1 y 30; cuando es mayor se utiliza directamente la tabla de la Normal como aproximación, ya que las diferencias que se obtienen son muy pequeñas.

Los programas estadísticos no necesitan utilizar esta aproximación Normal, ya que pueden calcular el valor exacto de *t* cualquiera que sea el número de grados de libertad. Las diferencias que se pueden observar entre los intervalos que se obtienen con STATISTICA o SPSS y la metodología que se explica a continuación se deben, por lo tanto, a que los programas estadísticos no tienen en cuenta si el número de datos es pequeño (<30) o grande (≥ 30), ya que siempre trabajan con la distribución *t* de Student y nunca utilizan la aproximación Normal.

Si el tamaño de la muestra es grande se usa la misma expresión descrita anteriormente para el caso de desviación típica conocida, pero con la diferencia de que se utiliza la cuasidesviación típica muestral (*s*) en vez de la desviación típica poblacional (σ).

$$I_{\mu}^{1-\alpha} = \left(\bar{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right)$$

Los valores críticos que se utilizan son las áreas bajo la curva Normal estándar (Tabla 1 del Apéndice I o archivo **Tabla 1.doc**).

En el Cuadro IV.1 se da un ejemplo de cómo estimar ese tipo de intervalo de confianza.

CUADRO IV.1. Intervalo de confianza de la media de una población Normal

EJEMPLO 1. Desviación típica conocida. Vamos a utilizar la misma serie de datos que en el Cuadro II.1 del Capítulo II (**Cuadro II.1.xls**). Se pretende calcular el intervalo de confianza al 95% de la media de esta serie de datos.

La media aritmética muestral y la desviación típica poblacional son 29,3 y 5,25, respectivamente.

El punto crítico de $z_{\frac{\alpha}{2}}$, considerando un α = 0,05, es decir $\frac{\alpha}{2} = 0{,}025$ y, por tanto, p = 0,975, lo consultamos en la Tabla 1 del Apéndice I (**Tabla 1.doc**). El valor de Z para este valor de p = 0,975 es de 1,96.

Por lo tanto, el intervalo de confianza al 95% es:

$$\left(29{,}3 - 1{,}96 * \frac{5{,}25}{\sqrt{20}}; 29{,}3 + 1{,}96\frac{5{,}25}{\sqrt{20}}\right)$$ Entre 26,99 y 31,6

EJEMPLO 2. Desviación típica desconocida para n ≥ 30. Supongamos que la concentración media en sangre de progesterona en un grupo de 42 mujeres es de 22,6 ngr ml^{-1} y la cuasidesviación típica muestral es de 1,2 ngr ml^{-1} ¿Cuál será el intervalo de confianza de la media al 98%?

Para un α = 0,02, es decir $\frac{\alpha}{2} = 0{,}01$ y, por tanto p = 0,99, $z_{\frac{\alpha}{2}}$ es aproximadamente igual a 2,33 (véase **Tabla 1.doc**)

Por tanto el intervalo de confianza de la media al 98% es:

$$\left(22{,}6 - 2{,}33 * \frac{1{,}2}{\sqrt{42}}; 22{,}6 + 2{,}33\frac{1{,}2}{\sqrt{42}}\right)$$ Entre 22,17 y 23,03

EJEMPLO 3. Desviación típica desconocida para n < 30. Supongamos que hemos calculado la estatura media de un grupo de 15 personas entre 20 y 25 años. La media fue 167 cm y la cuasidesviación típica 4,6 cm. Se quiere saber el intervalo de confianza al 98% de la media.

El punto crítico de $t_{\frac{\alpha}{2};14}$ considerando un α = 0,02, es decir $\frac{\alpha}{2} = 0{,}01$, es 2,624, lo consultamos en la Tabla 2 del Apéndice I (**Tabla 2.doc**).

El intervalo de confianza al 98% es:

$$\left(167 - 2{,}624 * \frac{4{,}6}{\sqrt{15}}; 167 + 2{,}624\frac{4{,}6}{\sqrt{15}}\right)$$ Entre 163,88 y 170,11

CUADRO IV.1. *(Continuación)*[R]

Aplicación con R.

EJEMPLO. En el archivo **Cuadro IV.1.V.csv** hay datos de altura y peso en hombres y mujeres de distintos países y ciudades. Se pretende calcular los intervalos de confianza al 95% de la altura y el peso, tanto para hombres como mujeres.

Paso 1. Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el *script* **Cuadro IV.1.R**, el cual es un archivo que contiene las instrucciones para calcular el intervalo de confianza de la media de una población Normal y también de una población que no tenga una distribución Normal, siempre que el tamaño de muestra sea mayor que 100, y la variable sea razonablemente simétrica. Para importar bases de datos desde Excel a R se recomienda guardar el archivo de Excel en formato ".csv" y luego leer desde R utilizando la función read.csv o read.csv2 según el tipo de formato ".csv" en el cual Excel haya guardado la respectiva base. Las instrucciones para realizar este tipo de importación de datos están en el Apéndice II.

Paso 2. La siguiente ventana muestra las instrucciones del *script* **Cuadro IV.1.R**. Lo único que hay que hacer, para una aplicación con otros datos, es cambiar la sección modificable por el usuario.

```
# Cuadro IV.1
# Intervalo de confianza de la media de una población normal
############################################################
# Sección modificable por el usuario
############################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro IV.1.V.csv",encoding="latin1")

# Selección de las variables de interés
#varInteres<-c("Altura.cm")
varInteres<-c("Altura.cm","Peso.Kg")

# Selección de categorización de las variables
# Si no se utilizan variables de agrupación obligatoriamente
# se exige que se coloce varAgrupación como NULL
#varAgrupacion<-NULL
varAgrupacion<-c("Sexo")
#varAgrupacion<-c("País","Ciudad","Sexo")

# Indique el nivel de significancia
alfa<-0.05

# Nombre del archivo de salida
ArchivodeSalida<-"Salida Cuadro IV.1.V.csv"

# ¿Los datos comprenden toda la población?
total<-"no"
```

CUADRO IV.1. *(Continuación)*[R]

Paso 3. En primer lugar hay que especificar el nombre del archivo donde están los datos.

```
datos<-read.csv2("Cuadro IV.1.V.csv",encoding="latin1")
```

Paso 4. A continuación hay que especificar a qué variables del archivo, se les quiere calcular los intervalos de confianza de la media. En este caso se utiliza tanto la altura como el peso.

```
#varInteres<-c("Altura.cm")
varInteres<-c("Altura.cm","Peso.Kg")
```

Paso 5. Es posible hacer los cálculos estadísticos agrupando las variables de la forma que se quiera. En el caso de seleccionar «NULL», no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna. En el ejemplo se selecciona que los cálculos se hagan a nivel de la variable «Sexo».

```
# varAgrupacion<-NULL
varAgrupacion<-c("Sexo")
#varAgrupacion<-c("País","Ciudad","Sexo")
```

Paso 6. A continuación se define el nivel de confianza.

```
alfa<-0.05
```

Paso 7. Es necesario especificar el nombre del archivo donde se guardarán todos los cálculos realizados.

```
ArchivodeSalida<-"Salida Cuadro IV.1.V.csv"
```

Paso 8. En la mayoría de los casos no es posible muestrear toda la población y, por tanto, se especificará «No» a la pregunta de si los datos comprenden toda la población.

```
total<-"no"
```

Paso 9. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*. Los resultados que se obtienen, y están en el archivo **Salida Cuadro IV.1.V.csv**, se muestran a continuación.

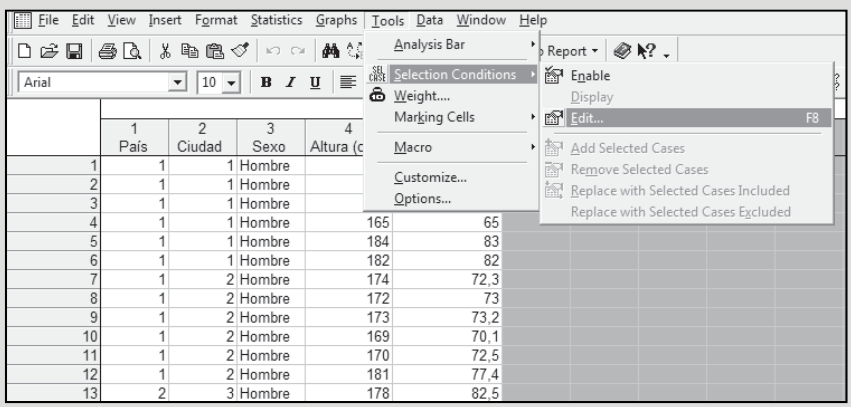
	A	B	C	D	E
1	Sexo	variable	LimInf	Media	LimSup
2	Hombre	Altura.cm	172,412561	175,703704	178,994846
3	Mujer	Altura.cm	160,332723	163,65	166,967277
4	Hombre	Peso.Kg	74,1519398	76,9407407	79,7295417
5	Mujer	Peso.Kg	58,2598145	60,35	62,4401855

CUADRO IV.1. *(Continuación)*[STATISTICA]

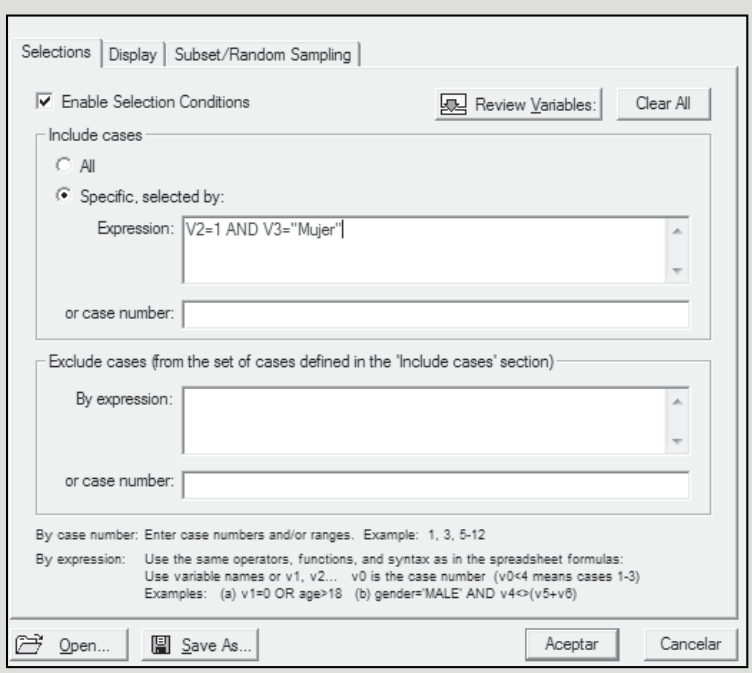
Aplicación con STATISTICA.

EJEMPLO. Vamos a utilizar los mismos datos que en el caso de R, los cuales están en el archivo **Cuadro IV.1.sta**. STATISTICA supone que solamente tenemos –como es habitual- datos de una muestra, por tanto sería el caso de desviación típica desconocida. Vamos a calcular el intervalo de confianza al 95% para la altura y el peso de las mujeres en la ciudad 1. Suponemos que ambas variables tienen distribución Normal.

Paso 1. Se selecciona «Selection Conditions» ⇨ «Edit».

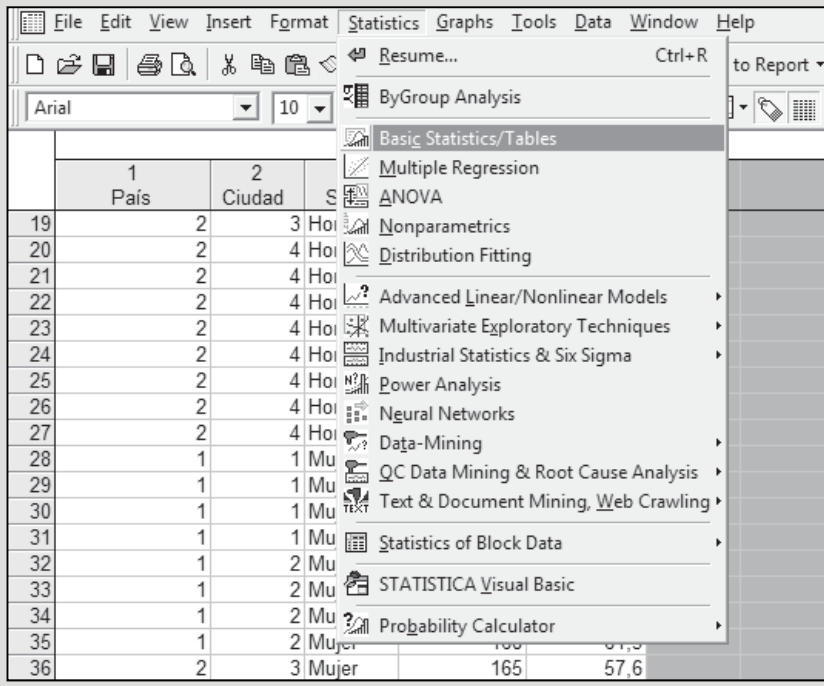


Paso 2. Se marca «Enable selection conditions» y «Specific.selected by:». En «Expresion» se introduce «V2=1 AND V3="Mujer"».

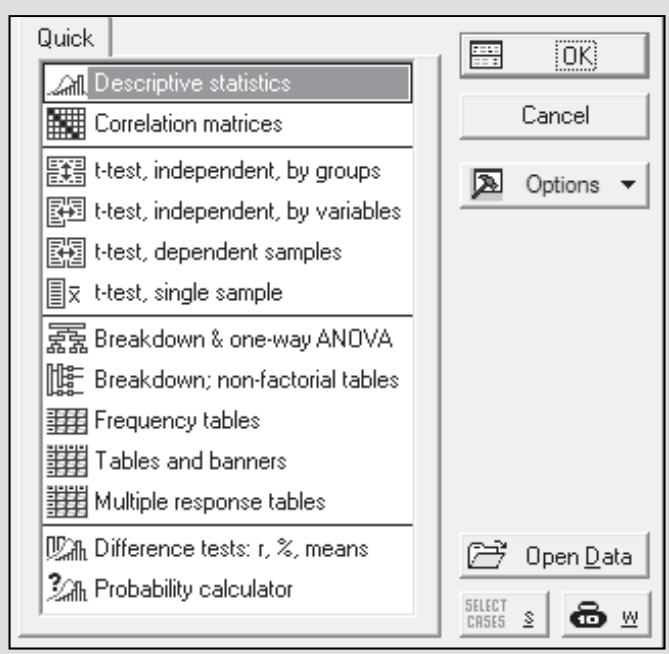

Selections
Display
Subset/Random Sampling
Enable Selection Conditions
Review Variables:
Clear All
Include cases
All
Specific, selected by:
Expression: V2=1 AND V3="Mujer"
or case number:
Exclude cases (from the set of cases defined in the 'Include cases' section)
By expression:
or case number:
By case number: Enter case numbers and/or ranges. Example: 1, 3, 5-12
By expression: Use the same operators, functions, and syntax as in the spreadsheet formulas:
Use variable names or v1, v2... v0 is the case number (v0<4 means cases 1-3)
Examples: (a) v1=0 OR age>18 (b) gender='MALE' AND v4<>(v5+v6)
Open...
Save As...
Aceptar
Cancelar

CUADRO IV.1. *(Continuación)*[STATISTICA]

Paso 3. Se selecciona «Statistics» ⇨ «Basic Statistics/Tables».

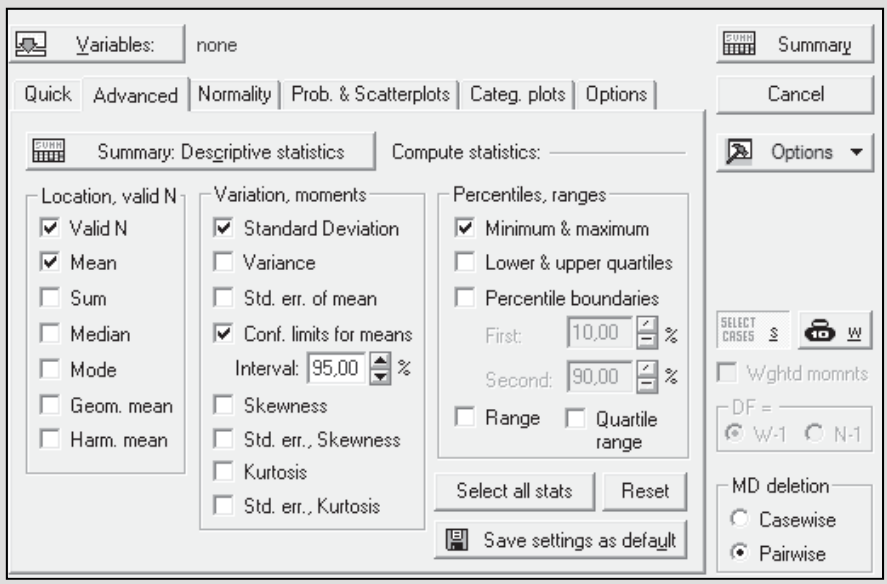


Paso 4. En la siguiente ventana seleccionamos «Descriptive statistics».

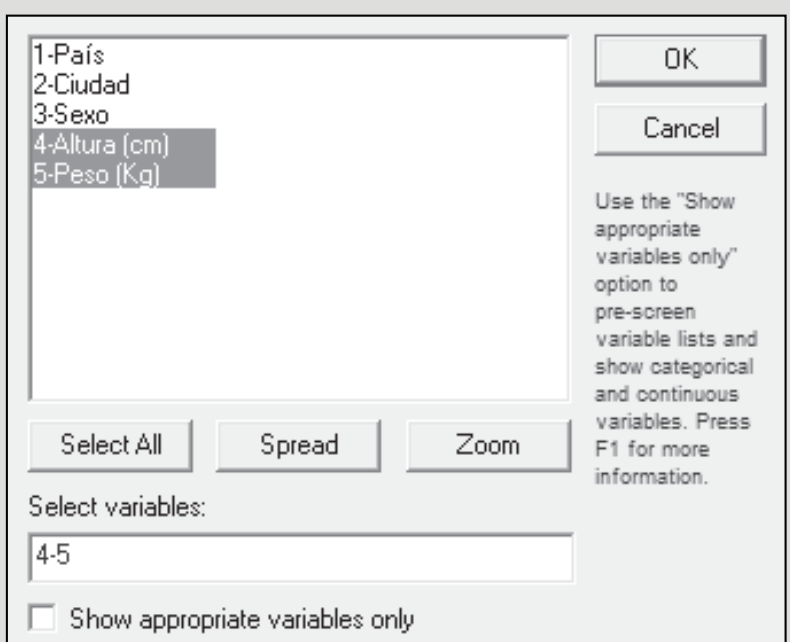


CUADRO IV.1. *(Continuación)* STATISTICA

Paso 5. A continuación en la ventana que aparece se selecciona la pestaña «Advanced» y se marca algunos estadísticos o se marca «Select all stats» para seleccionarlos todos.



Paso 6. En la ventana anterior pulsamos en «Variables» y en la ventana que aparece seleccionamos las variables que deseamos analizar.



CUADRO IV.1. *(Continuación)*[SPSS]

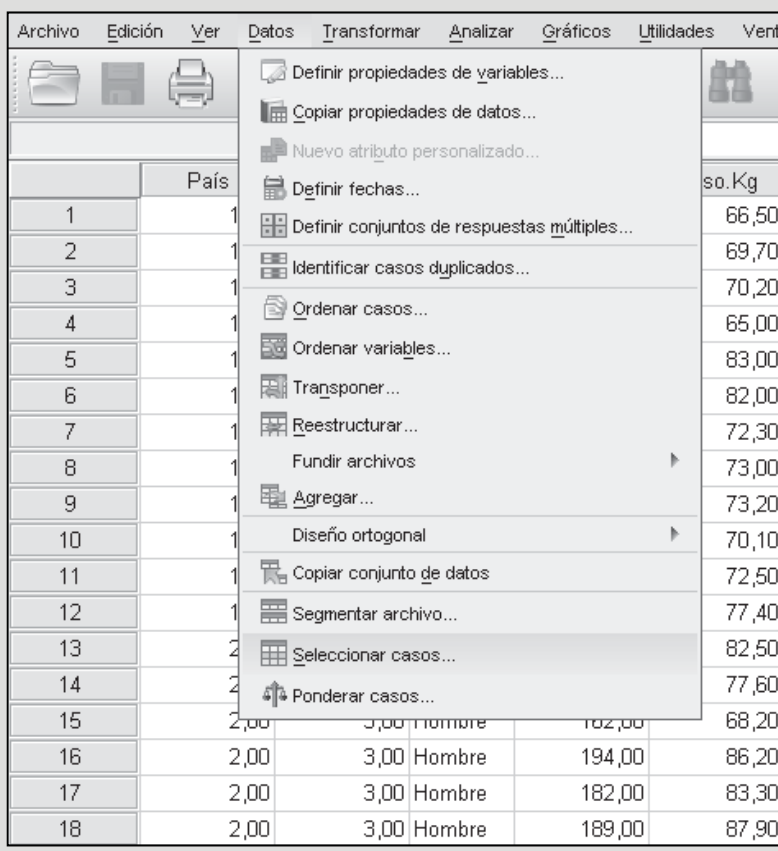
Paso 7. Por último, en la ventana del paso 5 pulsamos «Summary» y obtenemos los resultados.

Variable	Descriptive Statistics (Cuadro IV.1.sta) Include condition: V2=1 AND V3="Mujer"						
	Valid N	Mean	Confidence -95,000%	Confidence +95,000%	Minimum	Maximum	Std.Dev.
Altura (cm)	4	158,0000	145,4707	170,5293	147,0000	165,0000	7,874008
Peso (Kg)	4	58,4000	51,6278	65,1722	52,3000	62,2000	4,255976

Estimación con SPSS.

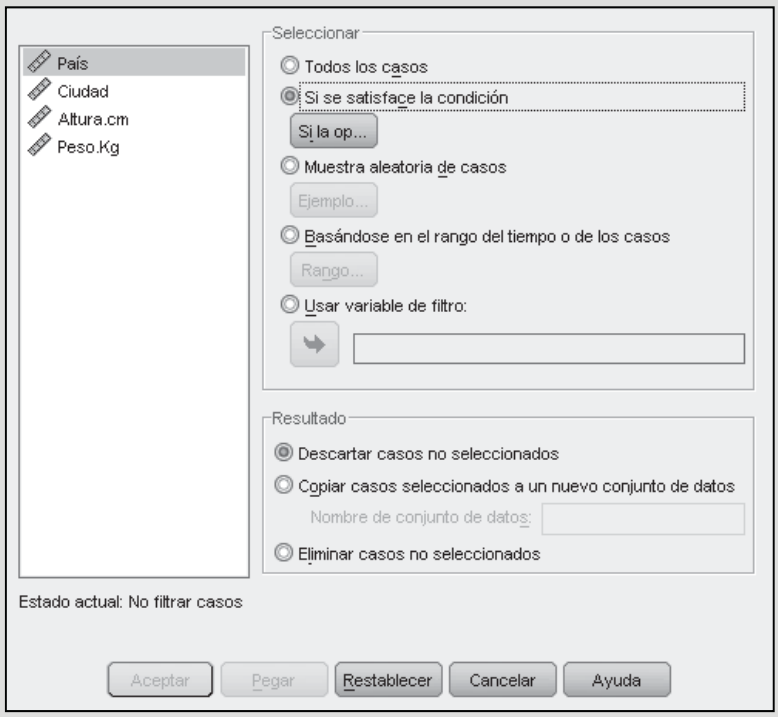
EJEMPLO. Se usarán los mismos datos que en la aplicación con R, los cuales están en el archivo **Cuadro IV.1.sav**. SPSS también supone que los datos se obtienen de una muestra y que las variables tienen distribución Normal, por tanto sería el caso de desviación típica desconocida. Vamos a calcular el intervalo de confianza al 95% para la altura y el peso de los hombres en la ciudad 2.

Paso 1. Para seleccionar los grupos que se van a analizar se entra en «Datos» ⇨ «Seleccionar casos…».

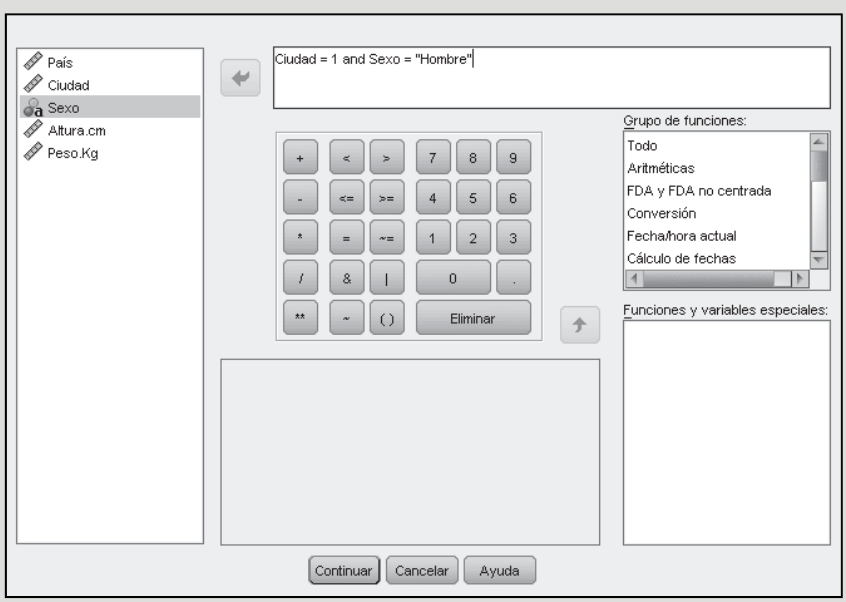


CUADRO IV.1. *(Continuación)*[SPSS]

Paso 2. Se marca «Si satisface la condición» y se pulsa «Si la op…».

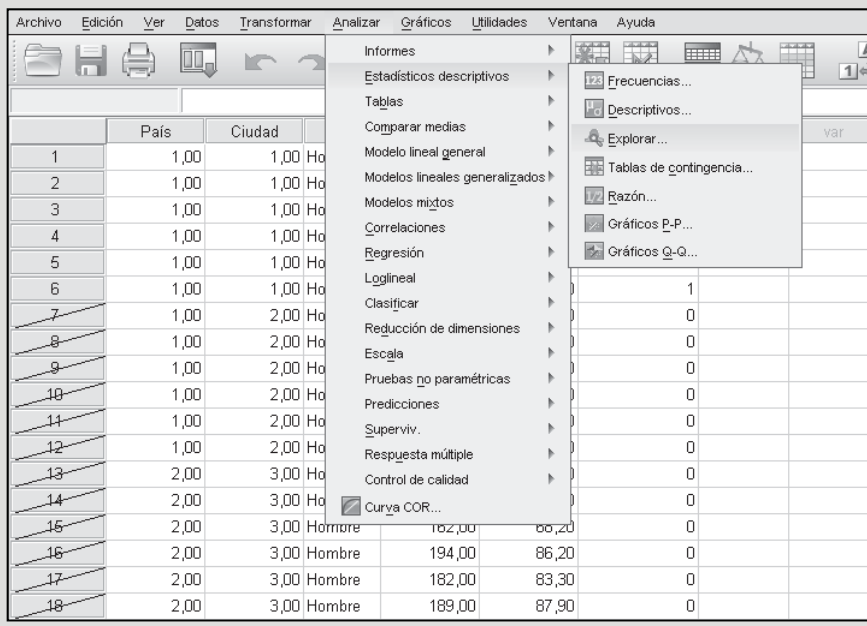


Paso 3. Se introduce la expresión «Ciudad = 1 and Sexo = "Hombre"». Se pulsa «Continuar» y «Aceptar » en la ventana del paso 2.

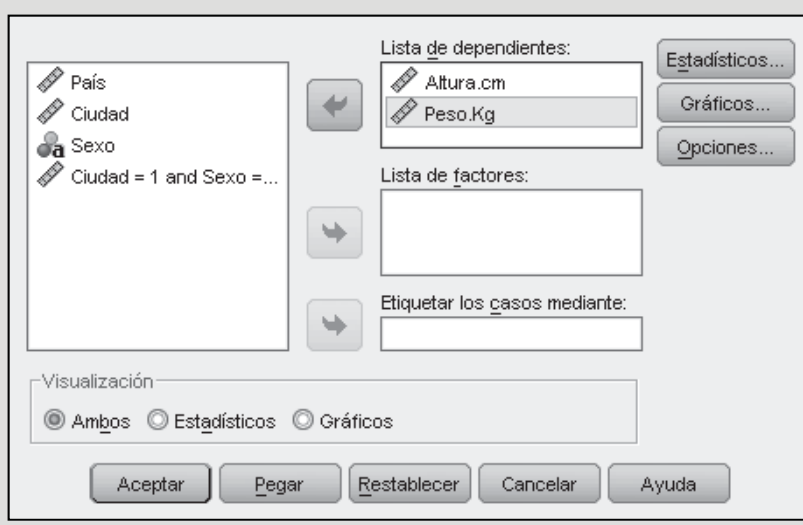


CUADRO IV.1. *(Continuación)*[SPSS]

Paso 4. Se entra en «Analizar» ⇨ «Estadísticos descriptivos» ⇨ «Explorar…».

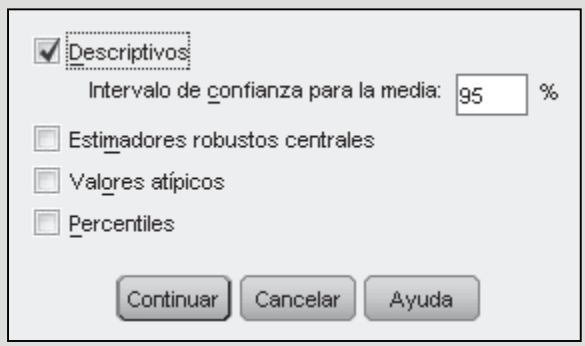


Paso 5. En la siguiente ventana se seleccionan las variables, que se introducen en la «Lista de dependientes». En el caso de querer realizar los cálculos agrupando los datos, se introducen en «Lista de factores» las variables que usamos para agrupar.



CUADRO IV.1. *(Continuación)*[SPSS]

Paso 6. En la ventana anterior en «Estadísticos» aparece la siguiente ventana donde seleccionamos «Descriptivos».



Paso 7. Los resultados se muestran a continuación.

Altura.cm	Media		173,3333	3,22146
	Intervalo de confianza para la media al 95%	Límite inferior	165,0523	
		Límite superior	181,6143	
	Media recortada al 5%		173,2037	
	Mediana		171,0000	
	Varianza		62,267	
	Desv. típ.		7,89092	
	Mínimo		165,00	
	Máximo		184,00	
	Rango		19,00	
	Amplitud intercuartil		16,00	
	Asimetría		,606	,845
	Curtosis		-1,710	1,741
Peso.Kg	Media		72,7333	3,19141
	Intervalo de confianza para la media al 95%	Límite inferior	64,5295	
		Límite superior	80,9371	
	Media recortada al 5%		72,5926	
	Mediana		69,9500	
	Varianza		61,111	
	Desv. típ.		7,81733	
	Mínimo		65,00	
	Máximo		83,00	
	Rango		18,00	
	Amplitud intercuartil		16,13	
	Asimetría		,709	,845
	Curtosis		-1,808	1,741

IV.2. Intervalo de confianza de la diferencia de medias de dos poblaciones Normales

En el Capítulo IX compararemos medias de distintas poblaciones. En este tipo de situaciones también es importante estimar el intervalo de confianza de la diferencia entre las medias. Al igual que con el intervalo anterior, el cálculo es diferente dependiendo de que se conozca la desviación típica poblacional (varianza conocida) o deba ser estimada con datos muestrales (varianza desconocida).

IV.2.1. VARIANZAS CONOCIDAS

El intervalo de confianza de μ_1 - μ_2 al nivel de confianza 1 - α ($I^{1-\alpha}_{\mu_1-\mu_2}$) se calcula por medio de la siguiente expresión:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left[\left(\bar{x}_1 - \bar{x}_2 \right) - z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \left(\bar{x}_1 - \bar{x}_2 \right) + z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

Los valores críticos que se utilizan son los correspondientes a la curva Normal estándar (Tabla 1 del Apéndice I o archivo **Tabla 1.doc**). En el Cuadro IV.2 se da un ejemplo de cómo estimar este intervalo de confianza.

IV.2.2. VARIANZAS DESCONOCIDAS

En el supuesto de que se quiera comparar dos medias, siendo las variables independientes y con distribución Normal, pero se desconocen las varianzas de la población, tenemos que aplicar otras expresiones que son diferentes dependiendo del tamaño de la muestra.

IV.2.2.1. Tamaño de muestra grande (≥ 30)

Para muestras grandes el intervalo de confianza de μ_1 - μ_2 al nivel de confianza 1 - α ($I^{1-\alpha}_{\mu_1-\mu_2}$) se calcula por medio de la siguiente expresión:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left[\left(\bar{x}_1 - \bar{x}_2 \right) - z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \left(\bar{x}_1 - \bar{x}_2 \right) + z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

Como vemos, la única diferencia con respecto a la situación en la que se conocían las varianzas, es que se usa la cuasivarianza en vez de la varianza.

Los valores críticos que se utilizan son los correspondientes a la curva Normal estándar (Tabla 1 del Apéndice I o archivo **Tabla 1.doc**). En el Cuadro IV.2 se da un ejemplo de cómo estimar este intervalo de confianza.

IV.2.2.2. Tamaño de muestra pequeño (< 30)

Se pueden dar dos situaciones: varianzas iguales o diferentes.

IV.2.2.2.1. Varianzas iguales

Suponiendo que desconocemos las varianzas, pero sabemos que no existen diferencias significativas entre ellas, en este caso se utiliza la distribución *t* de Student y el intervalo de confianza de μ_1 - μ_2 al nivel de confianza 1 - α ($I^{1-\alpha}_{\mu_1-\mu_2}$) se calcula por medio de la siguiente expresión:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left[\begin{array}{l} \left(\bar{x}_1 - \bar{x}_2\right) - t_{\frac{\alpha}{2};n_1+n_2-2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}, \\ \left(\bar{x}_1 - \bar{x}_2\right) + t_{\frac{\alpha}{2};n_1+n_2-2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \end{array} \right]$$

Se utilizan los valores críticos de la distribución *t* de Student que se pueden ver en la Tabla 2 del Apéndice I o archivo **Tabla 2.doc**. En el Cuadro IV.2 se da un ejemplo de cómo estimar este intervalo de confianza.

IV.2.2.2.2. Varianzas diferentes

En aquellos casos en los que se desconocen las varianzas y, además, existen diferencias significativas entre ellas, o simplemente no se sabe si las varianzas son iguales o no, el intervalo de confianza de μ_1 - μ_2 al nivel de confianza 1 - α ($I^{1-\alpha}_{\mu_1-\mu_2}$) se calcula de la siguiente forma:

$$I^{1-\alpha}_{\mu_1-\mu_2} = \left[\left(\bar{x}_1 - \bar{x}_2\right) - t_{\frac{\alpha}{2},f}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}, \left(\bar{x}_1 - \bar{x}_2\right) + t_{\frac{\alpha}{2},f}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \right]$$

donde *f* son los grados de libertad y se denomina aproximación de Welch, que se calcula del siguiente modo, tomando *f* el valor del entero más próximo:

$$f = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1+1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2+1}} - 2$$

Se utilizan los valores críticos de la distribución *t* de Student que se pueden ver en la Tabla 2 del Apéndice I o archivo **Tabla 2.doc**. En el Cuadro IV.2 se da un ejemplo de cómo estimar este intervalo de confianza.

CUADRO IV.2. Intervalo de la diferencia de medias de poblaciones Normales

En todos los ejemplos se parte de la base de que las poblaciones son Normales y las muestras independientes.

EJEMPLO 1. Varianzas conocidas. Una muestra de 55 hombres adultos del Amazonas presentaba una media de 157 cm en el año 1962. Otra muestra similar de 46 hombres en 2008 tiene una media de 162 cm. La varianza de las estaturas de toda la población de hombres adultos en el Amazonas era de 1,4 en 1962 y de 3,4 en 2008. ¿Cuál es el intervalo de confianza del aumento de la estatura, es decir de la diferencia de las medias, al 95%?

El punto crítico de $z_{\frac{\alpha}{2}}$, considerando un α = 0,05, es decir $\frac{\alpha}{2} = 0{,}025$ y, por tanto, p = 0,975, lo consultamos en la Tabla 1 del Apéndice I (**Tabla 1.doc**). El valor de Z para este valor de p = 0,975 es de 1,96.

Por lo tanto, el intervalo de confianza al 95% es:

$$\left((157-162)-1{,}96*\sqrt{\frac{1{,}4}{55}+\frac{3{,}4}{46}},(157-162)+1{,}96*\sqrt{\frac{1{,}4}{55}+\frac{3{,}4}{46}}\right)$$

Entre -5,62 y -4,38

EJEMPLO 2. Varianzas desconocidas para n ≥ 30. Los tamaños medios de la base del ala de dos especies de pájaros son 5,8 (47 individuos) y 4,9 cm (37 individuos), y las cuasivarianzas 1,2 y 0,9, respectivamente ¿Cuál es el intervalo de confianza de la diferencia de las medias al 98%?

Para un α = 0,02, es decir $\frac{\alpha}{2} = 0{,}01$ y, por tanto p = 0,99, $z_{\frac{\alpha}{2}}$ es aproximadamente igual a 2,33 (véase **Tabla 1.doc**).

Por tanto el intervalo de confianza de la diferencia de medias al 98% es:

$$\left((5{,}8-4{,}9)-2{,}33*\sqrt{\frac{1{,}2}{47}+\frac{0{,}9}{37}},(5{,}8-4{,}9)+2{,}33*\sqrt{\frac{1{,}2}{47}+\frac{0{,}9}{37}}\right)$$

Entre 0,38 y 1,42

EJEMPLO 3. Varianzas desconocidas y homogéneas para n < 30. La concentración media de vitamina E en 14 botellas de una marca de aceite de oliva fue de 0,22 mg ml^{-1} con una cuasivarianza de 0,05, y en 10 botellas de otra marca fue de 0,17 mg ml^{-1} con una cuasivarianza de 0,049. ¿Cuál es el intervalo de confianza de la diferencia de las medias al 98%?

CUADRO IV.2. *(Continuación)*

El punto crítico de $t_{\frac{\alpha}{2};22}$ considerando un α = 0,02, es decir $\frac{\alpha}{2} = 0{,}01$, lo consultamos en la Tabla 2 del Apéndice I (**Tabla 2.doc**) y es 2,508.

El intervalo de confianza al 98% es:

$$\left(\begin{array}{l} (0{,}22-0{,}17)-2{,}508*\sqrt{\frac{1}{14}+\frac{1}{10}}*\sqrt{\frac{(14-1)*0{,}05+(10-1)*0{,}049}{14+10-2}}, \\ (0{,}22-0{,}17)+2{,}508*\sqrt{\frac{1}{14}+\frac{1}{10}}*\sqrt{\frac{(14-1)*0{,}05+(10-1)*0{,}049}{14+10-2}} \end{array} \right)$$

Entre -0,181 y 0,281

EJEMPLO 4. Varianzas desconocidas y diferentes para n < 30. La concentración media de glóbulos rojos de 18 personas que habitan un pueblo a 1.200 metros de altura es de $4{,}85*10^6$ μl^{-1} con una cuasivarianza de 0,041, y de 22 personas que habitan un pueblo a nivel del mar es de $4{,}67*10^6$ μl^{-1} con una cuasivarianza de 0,015 ¿Cuál es el intervalo de confianza de la diferencia de las medias al 95%?

El valor de *f* es:

$$\frac{\left(\frac{0{,}041}{18}+\frac{0{,}015}{22}\right)^2}{\frac{\left(\frac{0{,}041}{18}\right)^2}{18+1}+\frac{\left(\frac{0{,}015}{22}\right)^2}{22+1}}-2=27{,}9$$

Por lo tanto, *f* = 28 grados de libertad. El punto crítico de $t_{\frac{\alpha}{2};28}$ considerando un α = 0,05, es decir $\frac{\alpha}{2} = 0{,}025$, lo consultamos en la Tabla 2 del Apéndice I (**Tabla 2.doc**) y es 2,048.

El intervalo de confianza al 95% es:

$$\left(\begin{array}{l} (4{,}85-4{,}67)-2{,}048*\sqrt{\frac{0{,}041}{18}+\frac{0{,}015}{22}}, \\ (4{,}85-4{,}67)+2{,}048*\sqrt{\frac{0{,}041}{18}+\frac{0{,}015}{22}} \end{array} \right)$$

Entre $0{,}07*10^6$ y $0{,}29*10^6$

CUADRO IV.2. *(Continuación)*[R]

Aplicación con R.

EJEMPLO. Vamos a utilizar los datos del archivo **Cuadro IV.1.V.csv**. El objetivo es estimar el intervalo de confianza de la diferencia de medias en el peso y en la altura, en hombres y mujeres, entre diferentes ciudades.

Paso 1. Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el *script* **Cuadro IV.2.R**, el cual es un archivo que contiene las instrucciones para calcular el intervalo de confianza de la diferencia de medias de poblaciones Normales.

Paso 2. La siguiente ventana muestra las instrucciones del *script* **Cuadro IV.2.R**. Lo único que hay que hacer, para una aplicación con otros datos, es cambiar la sección modificable por el usuario.

```
# Cuadro IV.2
# Intervalo de la diferencia de medias de poblaciones Normales

#########################################################
# Sección modificable por el usuario
#########################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro IV.1.V.csv",encoding="latin1")

# Selección de las variables de interés
#varInteres<-c("Altura.cm")
varInteres<-c("Altura.cm","Peso.Kg")

# Selección de categorización de las variables
# Si no se utilizan variables de agrupación obligatoriamente
# se exige que se coloce varAgrupación como NULL
#varAgrupacion<-NULL
#varAgrupacion<-c("Sexo")
varAgrupacion<-c("País","Sexo")

# Selección de variables con los niveles deseados de
# comparación.
# Si no se colocan niveles de comparación se supone
# que la variable es binaria.
#varSel<-list("Sexo")
#varSel<-list("Sexo",c("Ciudad","3","4"))
varSel<-list(c("Ciudad","1","2"))
```

Paso 3. En primer lugar se especifica el archivo donde están los datos.

datos<-read.csv2("Cuadro IV.1.V.csv",encoding="latin1")

Paso 4. A continuación hay que especificar a qué variables del archivo se les quiere calcular los intervalos de confianza de la diferencia de las medias. En este caso se utiliza tanto la altura como el peso.

#varInteres<-c("Altura.cm")
varInteres<-c("Altura.cm","Peso.Kg")

CUADRO IV.2. *(Continuación)*[R]

Paso 5. Es posible agrupar las variables de la forma que se quiera. En el caso de seleccionar «NULL», no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna. En el ejemplo se selecciona que los cálculos se hagan a nivel de «País» y «Sexo».

```
#varAgrupacion<-NULL
#varAgrupacion<-c("Sexo")
varAgrupacion<-c("País","Sexo")
```

Paso 6. A continuación se especifica la variable o variables que tienen los dos grupos a estudiar. Es importante mencionar que si la variable fue definida anteriormente como variable agrupación, no es posible usarla en esta sección. Si la variable tiene más de 2 grupos, es necesario especificar cuales son los grupos a estudiar. En el ejemplo se calcula el intervalo de confianza de la diferencia de las medias entre la ciudad 1 y la ciudad 2. Si la variable solo tiene 2 grupos, como por ejemplo «Sexo», no es necesario especificar los grupos a analizar.

```
#varSel<-list("Sexo")
#varSel<-list("Sexo",c("Ciudad","3","4"))
varSel<-list(c("Ciudad","1","2"))
```

Paso 7. A continuación se define el nivel de confianza.

```
alfa<-0.05
```

Paso 8. Es necesario especificar el nombre del archivo donde se guardarán todos los cálculos realizados.

```
ArchivodeSalida<-"Salida Cuadro IV.2.V.csv"
```

Paso 9. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*. Los resultados se muestran a continuación.

A	B	C	D	E	F	G	H	I	J	K	L	M
Tipo.de.Grupo	País	Sexo	variable	Freq	n1	n2	media1	media2	limInf	diferencia	limSup	valorPMedia
Ciudad.1_2	1	Hombre	Altura.cm	12	6	6	173,333	173,167	-7,991	0,1666667	8,3247	0,964588691
Ciudad.1_2	1	Mujer	Altura.cm	8	4	4	158	168,75	-22,16	-10,75	0,6562	0,060587882
Ciudad.1_2	1	Hombre	Peso.Kg	12	6	6	72,7333	73,0833	-8,54	-0,35	7,8404	0,91992711
Ciudad.1_2	1	Mujer	Peso.Kg	8	4	4	58,4	62,475	-9,894	-4,075	1,7441	0,137450766

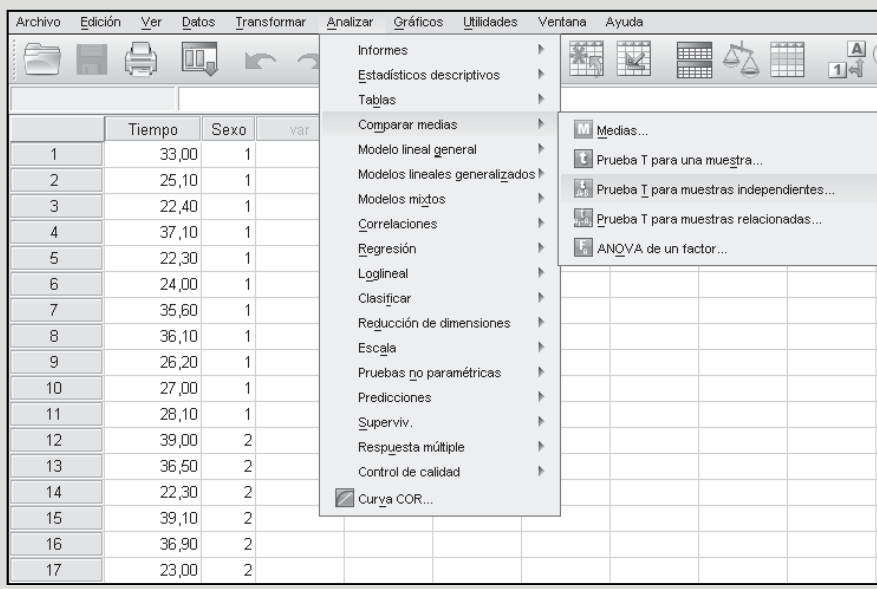
Por ejemplo, el intervalo de confianza de la diferencia en altura en los hombres entre la ciudad 1 y la 2 es entre -7,99 y 8,32, y no hay diferencias significativas entre las medias (p = 0,964), ni tampoco hay diferencias significativas entre las varianzas (p = 0,202). Por el contrario, al comparar el peso en los hombres entre la ciudad 1 y la 2, como el valor de probabilidad es menor que 0,05 (p = 0,021), hay diferencias significativas entre las varianzas, pero no entre las medias (p = 0,919).

CUADRO IV.2. *(Continuación)*[SPSS]

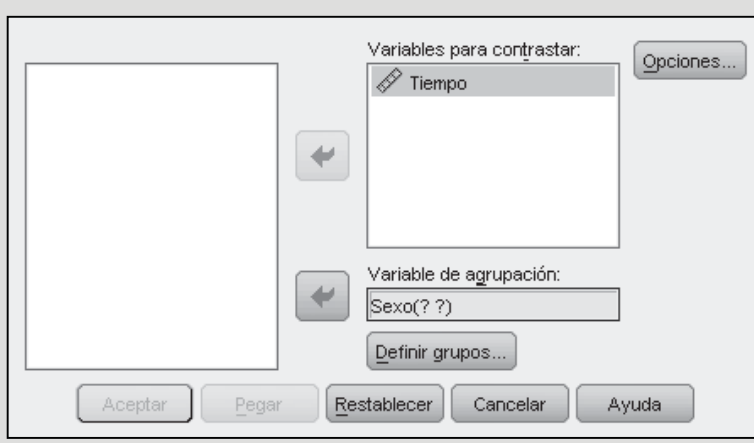
Aplicación con SPSS.

EJEMPLO. En el archivo **Cuadro IV.2.sav** hay datos del tiempo (en meses) que permanecen inmunes un grupo de mujeres y hombres después de que se les suministra una vacuna. Se quiere saber el intervalo de confianza al 95% de la diferencia de medias entre el tiempo que permanecen inmunes los hombres y las mujeres.

Paso 1. Se entra en «Analizar» ⇨ «Comparar medias» ⇨ «Prueba T para muestras independientes…»

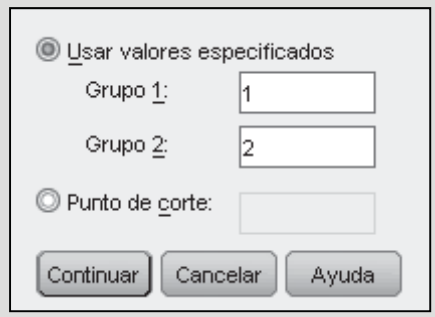


Paso 2. En la siguiente ventana se incluye «Tiempo» como variable dependiente y «Sexo» como factor.

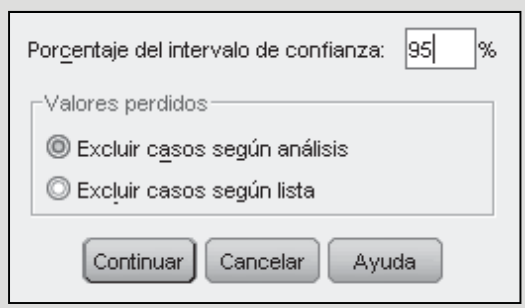


CUADRO IV.2. *(Continuación)*[SPSS]

Paso 3. En la ventana anterior, en «Definir grupos…», aparece la siguiente ventana donde se especifican los códigos de los grupos que se han usado; en este caso 1 para hombres y 2 para mujeres.



Paso 4. En la ventana del paso 2, en «Opciones», aparece la siguiente ventana donde se define el porcentaje del intervalo de confianza.



Paso 5. Los resultados se muestran en la siguiente ventana. Se observa que SPSS da el intervalo de confianza, tanto para el caso en el que se asumen varianzas iguales (entre -6,96 y 3,24), como para el caso de varianzas diferentes (entre -6,91 y 3,19).

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
									95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Inferior	Superior
Tiempo	Se han asumido varianzas iguales	,759	,393	-,756	22	,458	-1,86014	2,46092	-6,96377	3,24349
	No se han asumido varianzas iguales			-,763	21,920	,453	-1,86014	2,43664	-6,91449	3,19421

IV.3. Intervalo de confianza de la varianza y desviación típica de una población Normal

A veces es importante estudiar la mayor o menor concentración de valores alrededor de la media. Por ejemplo, cuando se estudia el tiempo que permanece un individuo inmune después de suministrarle una vacuna, no solo es importante conocer la duración media, sino también que la variabilidad de la duración del efecto no sea muy grande de unos individuos a otros. Para ello se puede estimar la varianza o la desviación típica mediante un intervalo de confianza.

El intervalo de la varianza al nivel de confianza 1 - α ($I_{\sigma^2}^{1-\alpha}$) se calcula por medio del intervalo estimador siguiente, donde s^2 es la cuasivarianza de la muestra:

$$I_{\sigma}^{1-\alpha} = \left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \right)$$

El intervalo de confianza de la desviación típica al nivel de confianza 1 - α ($I_{\sigma}^{1-\alpha}$) se estima por medio del intervalo estimador siguiente:

$$I_{\sigma}^{1-\alpha} = \left(\frac{\sqrt{n-1}s}{\sqrt{\chi^2_{\frac{\alpha}{2};n-1}}}, \frac{\sqrt{n-1}s}{\sqrt{\chi^2_{1-\frac{\alpha}{2};n-1}}} \right)$$

Se utilizan los valores críticos de la distribución χ^2 que se pueden ver en la Tabla 3 del Apéndice I o archivo **Tabla 3.doc**. En el Cuadro IV.3 se da un ejemplo de cómo estimar ese tipo de intervalo de confianza.

CUADRO IV.3. Intervalo de confianza de la varianza y desviación típica de una población Normal

EJEMPLO. Se usa el archivo **Cuadro IV.3.V.csv**, que tiene los datos del tiempo (en meses) que permanecen inmunes un grupo de mujeres y hombres después de que se les suministra una vacuna. Se quiere saber el intervalo de confianza al 95% de la desviación típica de las mujeres.

En la Tabla 3 del Apéndice I o archivo **Tabla 3.doc** se observa el valor:

$$\chi^2_{\frac{0,05}{2};12} = 23,337 \qquad \chi^2_{1-\frac{0,05}{2};12} = 4,404$$

CUADRO IV.3. (*Continuación*)^R

Por lo tanto, el intervalo de confianza al 95% de la desviación típica de las mujeres será:

$$\left(\frac{\sqrt{12}*6{,}31}{\sqrt{23{,}337}},\frac{\sqrt{12}*6{,}31}{\sqrt{4{,}404}}\right)$$ Entre 4,52 y 10,41

Aplicación con R.

EJEMPLO. Vamos a realizar con R el mismo caso que se explicó anteriormente. Los datos están en el archivo **Cuadro IV.3.V.csv**.

Paso 1. Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro IV.3.R**, el cual contiene las instrucciones para calcular el intervalo de confianza de la varianza y de la desviación típica de una población Normal.

Paso 2. La siguiente ventana muestra las instrucciones del archivo **Cuadro IV.3.R**. Lo único que hay que hacer, para una aplicación con otros datos, es cambiar la sección modificable por el usuario.

```
# Cuadro IV.3
# Intervalo de confianza de la varianza y desviación tipica

###########################################################
# Sección modificable por el usuario
###########################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro IV.3.V.csv",encoding="latin1")

# Selección de las variables de interés
varInteres<-c("Tiempo")

# Selección de categorización de las variables
# Si no se utilizan variables de agrupación obligatoriamente
# se exige que se coloce varAgrupación como NULL
# varAgrupacion<-NULL
varAgrupacion<-c("Sexo")


# Indique el nivel de significancia
alfa<-0.05

# Nombre del archivo de salida
ArchivodeSalida<-"Salida Cuadro IV.3.V.csv"

# Intervalo de confianza para varianza o desviacion tipica?
# tipoInterv<-"var"
tipoInterv<-"desv.est"
```

CUADRO IV.3. (*Continuación*)^{R}

Paso 3. En primer lugar hay que especificar el nombre del archivo donde están los datos.

```
datos<-read.csv2("Cuadro IV.3.V.csv",encoding="latin1")
```

Paso 4. A continuación hay que especificar las variables del archivo para las cuales se quiere calcular los intervalos de confianza. En este caso solo hay una variable, pero se pueden poner varias variables.

```
varInteres<-c("Tiempo")
```

Paso 5. Es posible hacer los cálculos estadísticos agrupando las variables de la forma que se quiera. En el caso de seleccionar «NULL», no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna. En el ejemplo se selecciona que los cálculos se hagan a nivel del sexo. Aunque en este ejemplo hay una sola variable de agrupación, es posible poner varias variables.

```
varAgrupacion<-c("Sexo")
```

Paso 6. A continuación se define el nivel de confianza.

```
alfa<-0.05
```

Paso 7. Es necesario especificar el nombre del archivo donde se guardarán todos los cálculos realizados.

```
ArchivodeSalida<-"Salida Cuadro IV.3.V.csv"
```

Paso 8. Por último, hay que especificar si se realiza el intervalo de confianza de la varianza o de la desviación típica. En este caso se selecciona el intervalo de la desviación típica.

```
# tipoInterv<-"var"
tipoInterv<-"desv.est"
```

Paso 9. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*. Los resultados que se obtienen, que están en el archivo **Salida Cuadro IV.3.V.csv**, se muestran a continuación.

	A	B	C	D	E
1	Sexo	variable	LimInf	DesviacionEstandar	LimSup
2	H	Tiempo	3,93079785	5,625736315	9,87279339
3	M	Tiempo	4,52279729	6,30718963	10,4114996
4					

IV.4. Intervalo de confianza de la razón de varianzas de dos poblaciones Normales

Determinar el intervalo de confianza de la razón de varianzas es importante, por ejemplo cuando queremos comparar la variabilidad de dos métodos a la hora de realizar un análisis bioquímico o la variabilidad de dos aparatos que nos dan el mismo tipo de medida. La comparación de varianzas entre las variables también es importante porque uno de los requisitos de los tests paramétricos (véase Capítulo IX) es que exista homogeneidad de varianzas entre las variables que se comparan.

El intervalo de la razón de las varianzas al nivel de confianza 1 - α ($I^{1-\alpha}_{\sigma_1^2/\sigma_2^2}$) se calcula de la siguiente forma:

$$I^{1-\alpha}_{\sigma_1^2/\sigma_2^2} = \left(\frac{s_1^2}{s_2^2} \frac{1}{F_{\frac{\alpha}{2};n_1-1,n_2-1}}, \frac{s_1^2}{s_2^2} F_{\frac{\alpha}{2};n_2-1,n_1-1} \right)$$

El intervalo de la razón de las desviaciones típicas al nivel de confianza 1 - α ($I^{1-\alpha}_{\sigma_1/\sigma_2}$) se calcula de la siguiente forma:

$$I^{1-\alpha}_{\sigma_1/\sigma_2} = \left(\frac{s_1}{s_2} \frac{1}{\sqrt{F_{\frac{\alpha}{2};n_1-1,n_2-1}}}, \frac{s_1}{s_2} \sqrt{F_{\frac{\alpha}{2};n_2-1,n_1-1}} \right)$$

Se utilizan los valores críticos de la distribución *F*, que se pueden ver en la Tabla 4 del Apéndice I o archivo **Tabla 4.doc**. En el Cuadro IV.4 se da un ejemplo de cómo estimar este tipo de intervalo de confianza para testar la homogeneidad de varianzas.

CUADRO IV.4. Intervalo de confianza de la razón de varianzas de poblaciones Normales y prueba de homogeneidad de varianzas entre poblaciones normales

EJEMPLO. Utilizando el mismo ejemplo del Cuadro IV.3 se quiere determinar si la variabilidad observada en hombres es igual a la observada en mujeres.

En la Tabla 4 del Apéndice I o archivo **Tabla 4.doc** se observa el valor de:

$$F_{\frac{0,05}{2};10,12} = 3,37 \text{ y } F_{\frac{0,05}{2};12,10} = 3,62$$

CUADRO IV.4. *(Continuación)*[R]

Por lo tanto, el intervalo de confianza al 95% de la razón de las desviaciones típicas será:

$$\left(\frac{5{,}62}{6{,}31}\frac{1}{\sqrt{3{,}37}}, \frac{5{,}62}{6{,}31}\sqrt{3{,}62}\right) \qquad \text{Entre 0,486 y 1,695}$$

Como el intervalo incluye el valor 1, que corresponde a dos desviaciones típicas iguales, podemos concluir que no hay diferencias significativas (con el nivel de confianza 95%) entre las desviaciones típicas de hombres y mujeres, es decir, la variabilidad en el tiempo que permanecen inmunes es la misma en hombres y en mujeres. En otras palabras, que existe homogeneidad de varianzas.

Aplicación con R.

EJEMPLO. El objetivo es estimar el intervalo de confianza de la razón de varianzas en el peso y en la altura, en hombres y mujeres, entre diferentes ciudades, y determinar si las varianzas son homogéneas. Los datos están en el archivo **Cuadro IV.1.V.csv**.

Paso 1. Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro IV.4.R**, el cual contiene las instrucciones para calcular el test de Levene y el test Brown. Ambos test permiten determinar si existe homogeneidad de varianzas entre dos series de datos. Para que este *script* funcione correctamente es necesario instalar el paquete "lawstat".

Paso 2. La siguiente ventana muestra las instrucciones del archivo **Cuadro IV.4.R**. Lo único que hay que hacer, para un aplicación con otros datos, es cambiar la sección modificable por el usuario.

```
# Cuadro IV.4
# Intervalo de confianza de la razón de varianzas
# de poblaciones normales y prueba de homogeneidad
# de varianza entre poblaciones normales

#########################################################
# Sección modificable por el usuario
#########################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro IV.1.V.csv",encoding="latin1")

# Selección de las variables de interés
#  varInteres<-c("Altura.cm")
varInteres<-c("Altura.cm","Peso.Kg")

# Selección de categorización de las variables
# Si no se utilizan variables de agrupación obligatoriamente
# se exige que se coloque varAgrupación como NULL
# varAgrupacion<-NULL
varAgrupacion<-c("Sexo")
#varAgrupacion<-c("País","Sexo")

# Selección de variables con los niveles deseados de
# comparación.
# Si no se colocan niveles de comparación se supone
# que la variable es binaria.
# varSel<-list("Sexo")
varSel<-list(c("Ciudad","1","2"),c("Ciudad","1","3"))
```

CUADRO IV.4. *(Continuación)*[R]

Paso 3. En primer lugar se especifica el archivo donde están los datos.

```
datos<-read.csv2("Cuadro IV.1.V.csv",encoding="latin1")
```

Paso 4. A continuación hay que especificar a qué variables del archivo se les quiere calcular los intervalos de confianza del cociente de varianzas. En este caso se utiliza tanto la «Altura» como el «Peso».

```
#varInteres<-c("Altura.cm")
varInteres<-c("Altura.cm","Peso.Kg")
```

Paso 5. Es posible agrupar las variables de la forma que se quiera. En el caso de seleccionar «NULL», no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna. En el ejemplo se selecciona que los cálculos se hagan a nivel de «País» y «Sexo».

```
#varAgrupacion<-NULL
#varAgrupacion<-c("Sexo")
varAgrupacion<-c("País","Sexo")
```

Paso 6. A continuación se especifica la variable o variables que tienen los dos grupos a estudiar. Es importante mencionar que si la variable fue definida anteriormente como variable agrupación, no es posible usarla en esta sección. Si la variable tiene más de 2 grupos, es necesario especificar cuales son los grupos a estudiar. En el ejemplo se calcula el intervalo de confianza del cociente de varianzas entre las ciudades 1 y 2, y entre las ciudades 1 y 3. Si la variable solo tiene 2 grupos, como por ejemplo «Sexo», no es necesario especificar los grupos a analizar.

```
# varSel<-list("Sexo")
varSel<-list(c("Ciudad","1","2"),c("Ciudad","1","3"))
```

Paso 7. A continuación se define el nivel de confianza.

```
alfa<-0.05
```

Paso 8. Es necesario especificar el nombre del archivo donde se guardarán todos los cálculos realizados.

```
ArchivodeSalida<-"Salida Cuadro IV.4.V.csv"
```

Paso 9. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*. Los resultados se muestran parcialmente a continuación, y están completos en el archivo **Salida Cuadro IV.4.V.csv**.

	A	B	C	D	E	F	G	H	I	J	K
1	Tipo.de.Grupo	Sexo	variable	Freq	n1	n2	desv.Est1	desv.Est2	limInf	razon	limSup
2	Ciudad.1_2	Hombre	Altura.cm	27	6	6	7,891	4,262	0,480	3,428	24,494
3	Ciudad.1_2	Mujer	Altura.cm	20	4	4	7,874	4,992	0,161	2,488	38,417
4	Ciudad.1_2	Hombre	Peso.Kg	27	6	6	7,817	2,388	1,500	10,718	76,595

CUADRO IV.4. *(Continuación)*[STATISTICA]

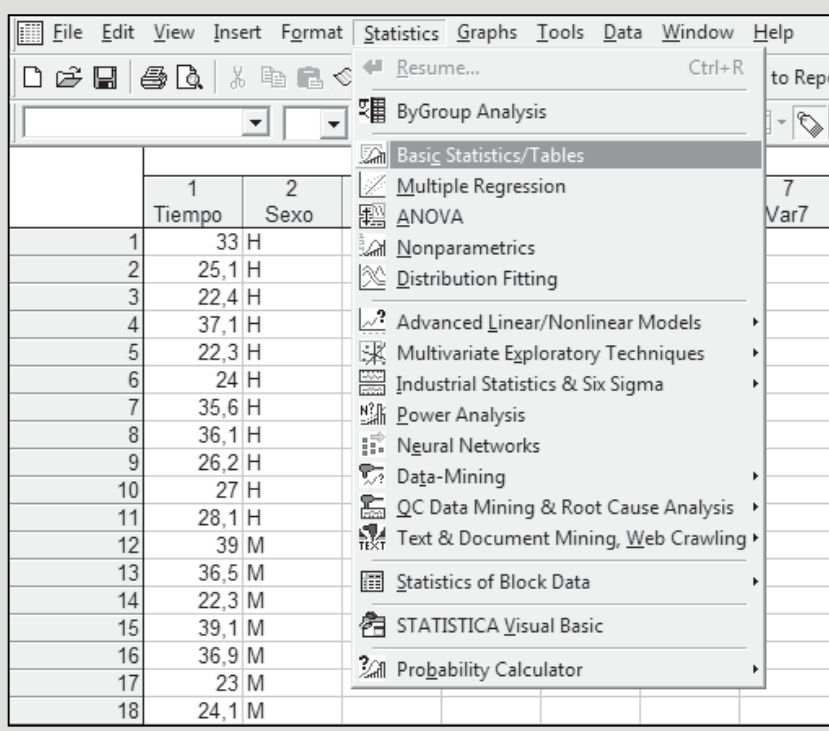
Por ejemplo, en el caso de la comparación de la altura de los hombres entre las ciudades 1 y 2, el intervalo de confianza de la razón de varianzas es entre 0,479 y 24,494. Tanto la probabilidad de la razón de varianzas (p = 0,202), como la del test de Levene (p = 0,083) y la del test Brown & Forsythe (p = 0,217) son mayores de 0,05, por lo tanto se acepta la hipótesis nula de que no existen diferencias significativas en las varianzas de la altura de los hombres entre las ciudades 1 y 2, es decir, que existe homogeneidad de varianzas.

Sin embargo, al comparar el peso de los hombres entre la ciudades 1 y 2, la probabilidad de la razón de varianzas (p = 0,021) y la del test de Levene (p = 0,007) son menores de 0,05, aunque no la del test Brown & Forsythe (p = 0,112), por lo tanto, es razonable pensar que se rechaza la hipótesis nula y, por tanto, existen diferencias significativas entre las varianzas de la altura de los hombres entre las ciudades 1 y 2, es decir, que no existe homogeneidad de varianzas.

Aplicación con STATISTICA.

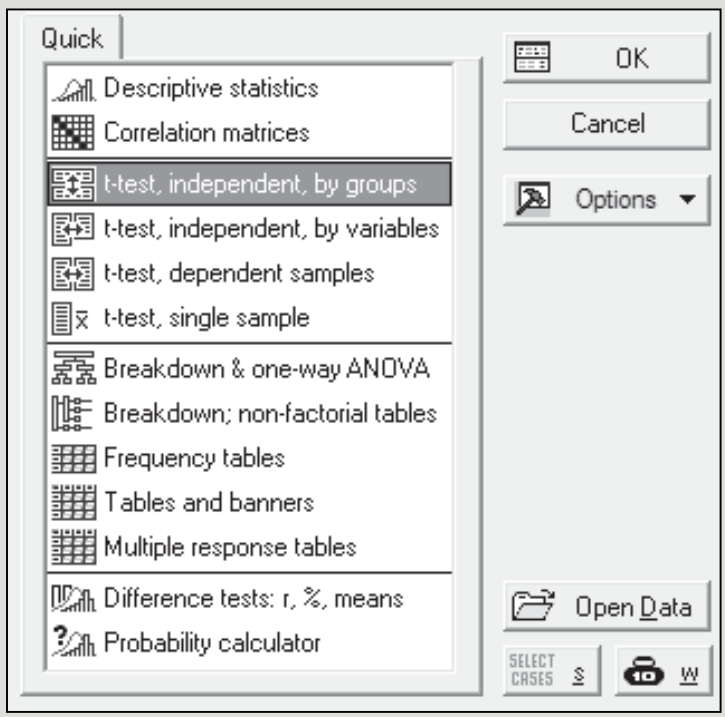
EJEMPLO. El objetivo es determinar si existen diferencias significativas en el tiempo que permanecen hombres y mujeres inmunes a una vacuna, y determinar si existe homogeneidad de varianzas. Los datos están en el archivo **Cuadro IV.4.sta**.

Paso 1. Se selecciona «Statistics» ⇨ «Basic Statistics/Tables».

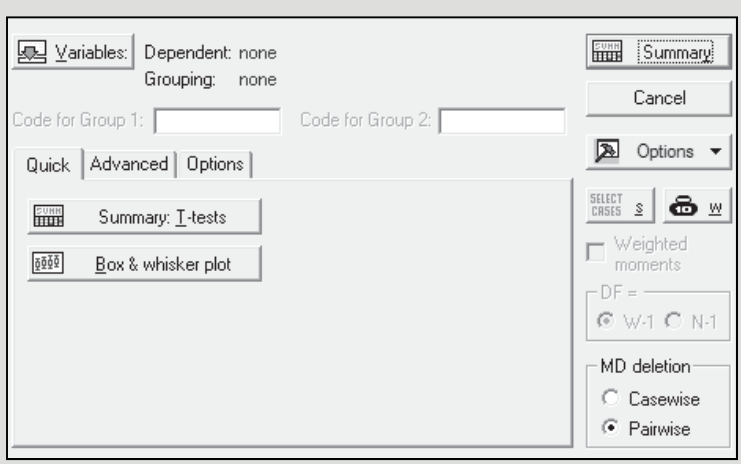


CUADRO IV.4. *(Continuación)*[STATISTICA]

Paso 2. Nos aparece la siguiente ventana donde podemos entrar por dos sitios diferentes dependiendo de cómo tengamos organizados los datos. Si tenemos dos variables en dos columnas, entonces entramos en «t-test, independent, by variables». Si los datos están organizados de tal forma que ambas variables están en una columna y en una segunda hay un código que identifica los grupos, como es el caso en este ejemplo, entonces se entraría en «t-test, independent, by groups».

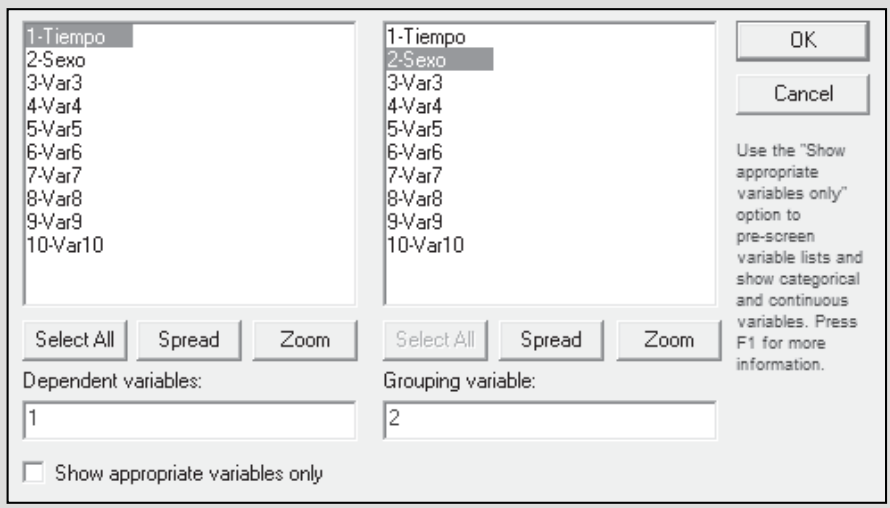


Paso 3. En la ventana siguiente entramos en «Variables» para especificar las variables.

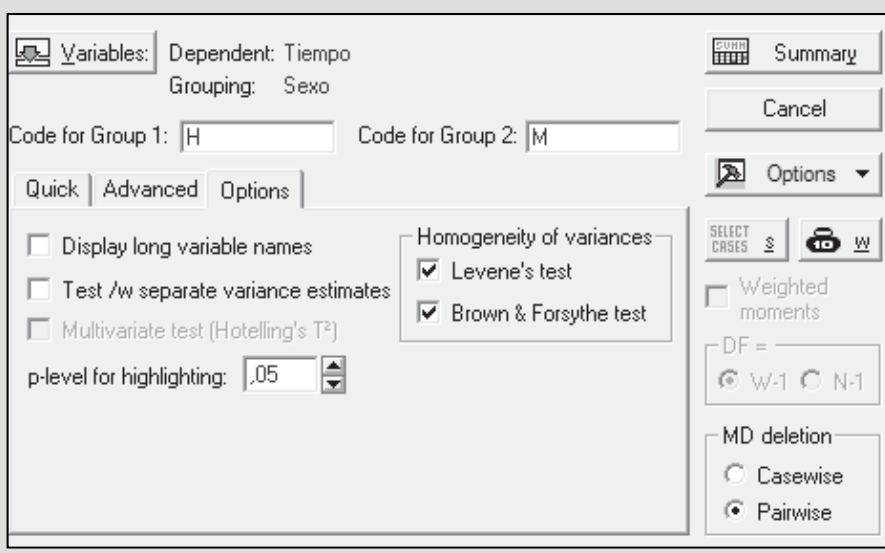


CUADRO IV.4. *(Continuación)*[STATISTICA]

Paso 4. En la ventana que nos sale especificamos la variable que se va a analizar («Dependent variables») y la que actúa como variable de agrupación («Grouping variable»).



Paso 5. Por último, en la pestaña «Options» marcamos en «Levene´s test» y en «Brown & Forsythe test» y pulsamos «Summary»



CUADRO IV.4. *(Continuación)*[SPSS]

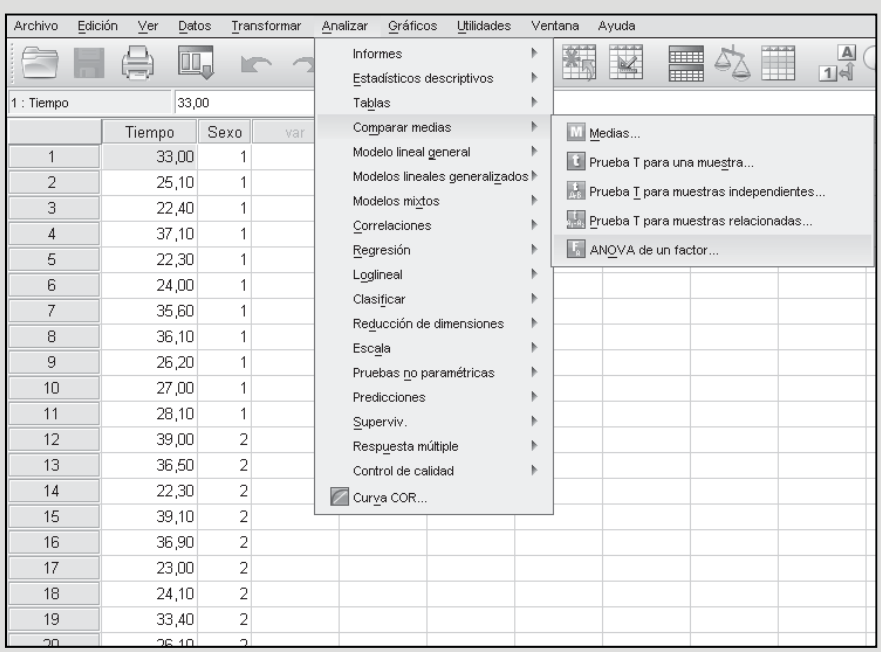
Paso 6. Los resultados del test de Levene y del test Brown-Forsythe se muestran en la siguiente tabla. Tanto la probabilidad del test de Levene (p = 0,393) como la del test Brown & Forsythe (p = 0,442) son mayores de 0,05, por lo tanto se acepta la hipótesis nula de que no existen diferencias significativas entre las varianzas de hombres y mujeres, es decir, que existe homogeneidad de varianzas.

	T-tests; Grouping: Sexo (Cuadro IV.4.sta) Group 1: H Group 2: M					
Variable	Levene F(1,df)	df Levene	p Levene	Brn-Fors F(1,df)	df Brn-Fors	p Brn-Fors
Tiempo	0,758639	22	0,393160	0,613888	22	0,441682

Aplicación con SPSS.

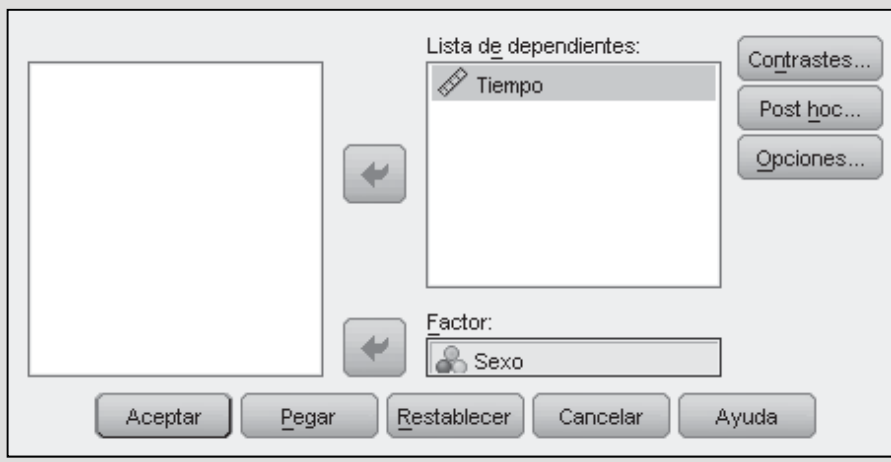
EJEMPLO. Al igual que en el caso con STATISTICA, el objetivo es determinar si existen diferencias significativas en la variabilidad del tiempo que permanecen hombres y mujeres inmunes a una vacuna, es decir, si existe homogeneidad de varianzas. Los datos están en el archivo **Cuadro IV.4.sav**.

Paso 1. Se entra en «Analizar» ⇨ «Comparar medias» ⇨ «ANOVA de un factor…»

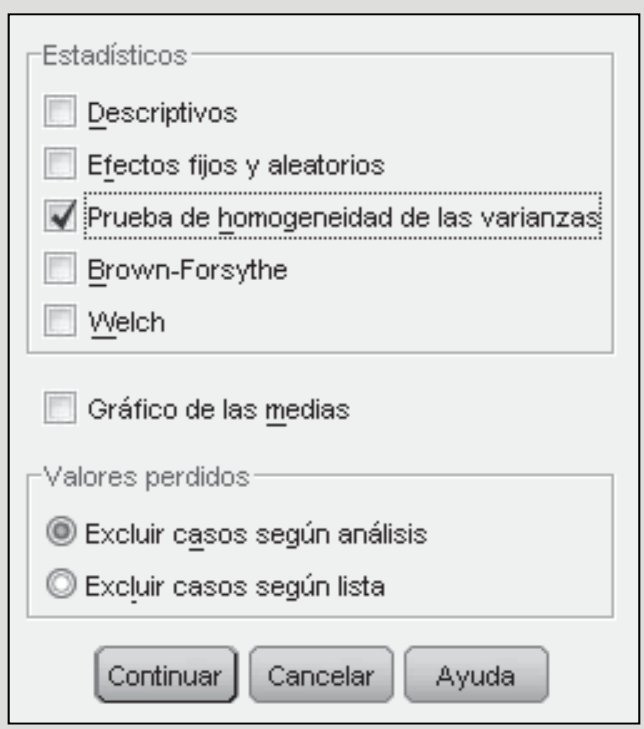


CUADRO IV.4. *(Continuación)*[SPSS]

Paso 2. En la siguiente ventana se incluye «Tiempo» como variable dependiente y «Sexo» como factor. Posteriormente se pulsa en «Opciones…»



Paso 3. En la siguiente ventana se marca «Prueba de homogeneidad de las varianzas».



CUADRO IV.4. *(Continuación)*[SPSS]

Paso 4. Los resultados se muestran en la siguiente tabla. Debido a que la probabilidad del test de Levene (p = 0,393) es mayor que 0,05, se acepta la hipótesis nula de que no existen diferencias significativas entre las varianzas de hombres y mujeres, es decir, que existe homogeneidad de varianzas.

Prueba de homogeneidad de varianzas

Tiempo

Estadístico de Levene	gl1	gl2	Sig.
,759	1	22	,393

Capítulo V

CONTRASTE DE HIPÓTESIS

Uno de los aspectos más importantes en cualquier tratamiento de datos es el *contraste de hipótesis*, que comenzaremos a emplear en los capítulos siguientes. El contraste de hipótesis se define como el procedimiento estadístico que permite determinar la verdad o falsedad de una afirmación acerca de uno o más parámetros. Por ejemplo, cuando nos planteamos si existen diferencias entre las medias de dos variables, tenemos que proponer una hipótesis de partida y, por medio del método estadístico más apropiado, concluir en términos probabilísticos, si la hipótesis inicial se acepta o se rechaza. En este capítulo vamos a dar una serie de nociones básicas sobre el contraste de hipótesis necesarias para entender los planteamientos de problemas basados en el contraste que veremos en los capítulos que vienen a continuación.

V.1.Tipos de hipótesis

Como se mencionó anteriormente, la base de los planteamientos de contraste es formular una hipótesis y determinar si es verdadera o falsa. Existen dos tipos de hipótesis:

1. La hipótesis nula (H_0). Es la hipótesis básica que se formula y se quiere contrastar y, por tanto, es la hipótesis que se acepta o se rechaza.
2. La hipótesis alternativa (H_a). Es distinta de H_0 e incompatible con ella. Puede haber varias hipótesis alternativas, y se elige la más adecuada a partir de la información disponible.

V.2. Estadístico de contraste

Una vez planteada la hipótesis tenemos que elegir el estadístico de contraste más apropiado, el cual es una variable aleatoria que seguirá una función de probabilidad y para cada muestra de datos tomará un determinado valor, que al compararlo con los valores críticos de esa función de probabilidad, nos permitirá aceptar o rechazar la hipótesis nula.

V.3. Tipos de errores

Existen dos tipos de errores:

1. El *error tipo I* es el que cometemos cuando rechazamos la hipótesis nula siendo verdadera. Su probabilidad se representa generalmente por α y se conoce como *nivel de significación*. El valor máximo que se le suele dar a α es 0,05, lo cual significa que rechazamos un 5% de las veces la hipótesis nula siendo cierta. Se puede fijar un valor de α más bajo, como por ejem-

plo 0,01, pero el problema es que aumentamos la probabilidad de cometer el siguiente tipo de error.

2. El *error tipo II* es el que cometemos cuando aceptamos la hipótesis nula siendo falsa. Su probabilidad se representa por β. Este error se tiene en cuenta determinando el tamaño de muestra necesario para garantizar el valor de β prefijado. En el supuesto de que no se tenga en cuenta el tamaño de muestra necesario y, por tanto, se omita el error de tipo II, entonces el procedimiento se suele denominar *contraste de significación*, ya que solo tiene en cuenta el error de tipo I.

El error de tipo II no se suele tener en cuenta porque, normalmente, se desconoce la información necesaria para ello.

V.4. Regiones críticas y de aceptación

El error de tipo I (probabilidad α) nos define la región de valores del estadístico que se denomina *región crítica* (Figura V.1). Cuando el valor del estadístico de contraste que hemos utilizado está dentro de la región crítica se rechaza la hipótesis nula.

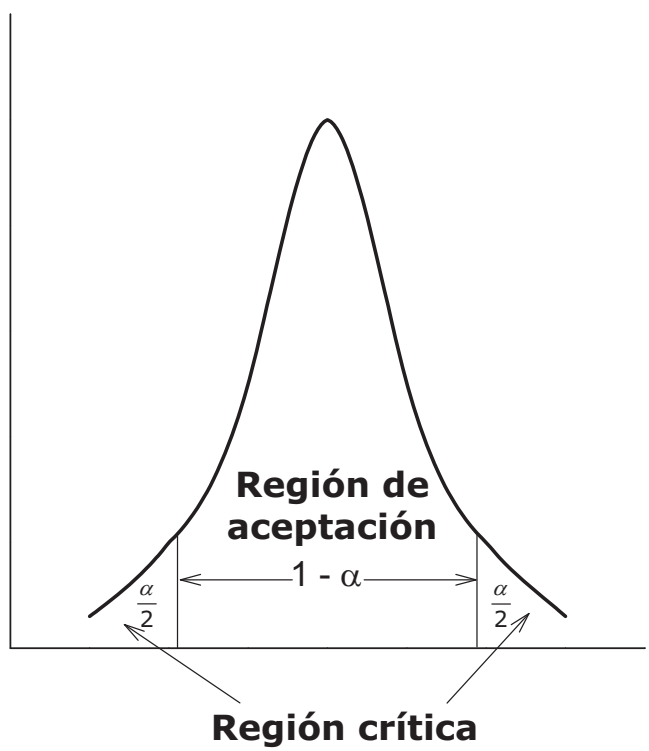


Figura V.1. Representación de las regiones de aceptación y crítica en la función de probabilidad de un estadístico de contraste.

La región complementaria recibe el nombre de *región de aceptación* (Figura V.1). Cuando el valor del estadístico de contraste que hemos utilizado está dentro de la región de aceptación, lo que ocurrirá con probabilidad 1 - α si H_0 es cierta, se acepta la hipótesis nula.

Cuando el valor de probabilidad es menor de 0,05 rechazamos la hipótesis nula H_0. Cuando no lo es, ¿significa eso que H_0 es cierta? No necesariamente. A menudo ocurre que la muestra es demasiado pequeña para rechazar H_0 pero con una muestra más grande obtendríamos otro valor de probabilidad que nos haría rechazarla. Se suele decir que "no rechazamos H_0", más correcto que "aceptamos H_0". El rechazo de H_0 con una probabilidad muy pequeña es concluyente, pero la aceptación –cuando la probabilidad es mayor que el nivel de significación (normalmente 0,05)– no lo es.

Por otra parte, con los mismos datos podríamos en muchos casos plantear distintas H_0 incompatibles entre sí, que podrían todas ellas ser "no rechazadas", aunque obviamente no podrían ser "aceptadas" al mismo tiempo por ser incompatibles. Esto ocurre por ejemplo con el contraste Kolmogorov-Smirnov de bondad de ajuste: podemos "no rechazar" la hipótesis nula de que la variable tiene distribución Normal y al mismo tiempo "no rechazar" que tiene distribución exponencial (por lo tanto en ningún caso estamos demostrando que la distribución es Normal).

Para establecer correctamente la región crítica es necesario tener en cuenta la hipótesis alternativa, y el tipo de comparación que se realiza:

1. En el *contraste bilateral*, la hipótesis nula es que un determinado parámetro tiene un valor específico, mientras que en la hipótesis alternativa el valor del parámetro es distinto. Por ejemplo, H_0: $\mu = \mu_1$, frente a H_a: $\mu \neq \mu_1$. En este caso la región crítica estaría formada por las dos colas.
2. En el *contraste unilateral*, la hipótesis nula es que un determinado parámetro tiene un valor específico H_0: $\mu = \mu_1$, frente a la hipótesis alternativa de que es mayor H_a: $\mu > \mu_1$ (la región crítica es la cola de la derecha) o que es menor H_a: $\mu < \mu_1$ (la región crítica es la cola de la izquierda).

Para un determinado contraste ¿debemos utilizar el valor de probabilidad unilateral o bilateral? Eso depende de la información que tengamos a priori para establecer la hipótesis alternativa. Por ejemplo, en un análisis de correlación, si tenemos evidencias previas o consideraciones teóricas que indican que debe existir una correlación positiva, entonces H_0 es R = 0 frente a H_a: R > 0, y la región crítica es unilateral. Por el contrario, si no tenemos información previa que indique el sentido de la correlación, H_0 sigue siendo R = 0, pero la alternativa H_a es "R distinto de cero", y el contraste es bilateral.

Tener más información nos permite utilizar un contraste unilateral, siempre preferible al bilateral ya que el valor de probabilidad es más pequeño (en torno a la mitad, o exactamente la mitad con distribuciones simétricas del estadístico) y el contraste es más sensible, porque encuentra diferencias significativas con muestras más pequeñas. Si tenemos dudas acerca de si tenemos suficiente información a priori para justificar un contraste unilateral, es más prudente en general utilizar la opción bilateral; por otra parte, si encontramos una diferencia significativa (si se rechaza H_0) con un contraste bilateral, también será significativa con un contraste unilateral (al revés no es necesariamente cierto).

V.5. Tipos de contrastes

La mayoría de los contrastes de hipótesis pueden agruparse en alguno de los siguientes tipos:

1. El contraste de bondad de ajuste consiste en el planteamiento de hasta qué punto una muestra se puede considerar como perteneciente a una población con una distribución teórica ya conocida. Es un método que se utiliza frecuentemente para determinar si una serie de datos presenta una distribución Normal, de Poisson, etc. (véase Capítulo VI).
2. El contraste de independencia o asociación determina si dos caracteres *X* e *Y* de una población son dependientes o independientes (véase Capítulo VIII para variables cualitativas y Capítulo X para variables cuantitativas). Por ejemplo, queremos determinar si la presencia de pesticidas en las personas tiene alguna relación con la zona geográfica.
3. El contraste de homogeneidad permite determinar si varias muestras tomadas de la misma población o de poblaciones diferentes se diferencian en un determinado carácter A (véase Capítulo VII para variables cualitativas y Capítulo IX para variables cuantitativas). Por ejemplo, hemos seleccionado varios grupos de individuos de una población a los que hemos sometido a la misma dosis de diferentes metales y queremos determinar si los metales afectan de forma diferente a la supervivencia de los individuos.

Capítulo VI

CONTRASTES DE BONDAD DE AJUSTE

Para determinar si las variables se ajustan a alguna de las distribuciones descritas en el Capítulo III, es decir, si se ajustan a una distribución Normal, Poisson, Binomial, etc., es necesario utilizar contrastes estadísticos, los cuales se denominan pruebas de bondad de ajuste.

Existen distintas pruebas de bondad de ajuste que se utilizan en función del tipo de datos y la distribución teórica esperada. Una clasificación de los ajustes más empleados es:

1. Muestras categorizadas (los datos son códigos asignados a los valores de una variable cualitativa o a clases en las que se agrupan los valores de una variable cuantitativa).
    - Chi-cuadrado
2. Muestras no categorizadas (variables cuantitativas, continuas o discretas, no agrupadas en intervalos o clases).
    - Para todas las distribuciones.
      Prueba Kolmogorov-Smirnov (test *K-S*).
    - Distribución Normal.
      Contraste de normalidad Shapiro-Wilk.

VI.1. Chi-cuadrado

Se puede aplicar tanto a distribuciones continuas (con los datos previamente agrupados en clases) como a distribuciones discretas o variables cualitativas. Se basa en cuantificar las diferencias entre las frecuencias observadas en cada clase y las esperadas, partiendo de la hipótesis nula de que los datos se ajustan a una distribución *F(x)* (que puede ser Normal, Poisson, etc.). Para su aplicación, en las *n* clases existentes, se contabiliza el número de casos observados (O_i) y, a través de la función teórica que se desea testar, se calcula el número de casos esperados (E_i). A partir de estas frecuencias se calcula el valor del estadístico χ^2, con la siguiente fórmula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Se aconseja que si el número de casos esperados dentro de una clase es menor de 5 se combinen varias clases en una, hasta conseguir que todas tengan una frecuencia esperada mayor o igual a 5. No se debe utilizar cuando hay pocos datos.

A continuación se deben calcular los grados de libertad (υ) de la muestra. Si los valores esperados se pueden calcular de forma previa al muestreo, el número de υ es *n* - 1. Sin embargo, si para calcular los valores esperados es necesario estimar algunos parámetros mediante estadísticos obtenidos en el muestreo (como, por ejemplo, μ o σ para una distribución Normal) el número de υ es *n* - *r* -1 donde *r* es el número de estadísticos necesarios para calcular el valor esperado. Una vez determinados χ^2 y υ, se busca en las tabla χ^2 (Tabla 3 en el Apéndice I o archivo **Tabla 3.doc**) el valor crítico para esos υ y para el nivel de significación elegido (generalmente α = 0,05). Si el χ^2 crítico es mayor que χ^2 calculado se acepta la hipótesis nula de que los datos observados se ajustan a la distribución esperada, mientras que si χ^2 crítico es menor que χ^2 calculado la hipótesis se rechaza. La mayoría de los programas estadísticos calculan directamente el valor *p* del contraste; si es menor que el nivel de significación elegido (0,05 o 0,01 habitualmente) la hipótesis se rechaza. En el Cuadro VI.1 se muestra cómo realizar este test.

VI.2. Test de Kolmogorov-Smirnov

Es la prueba adecuada para testar la normalidad de una muestra si el número de datos es grande (*n* > 30), aunque se puede usar tanto para muestras grandes como pequeñas. También se puede usar para testar otras distribuciones como la Binomial o de Poisson. Es un test muy conservador que se aplica a variables continuas. Se basa en la determinación de la máxima diferencia (*D*) entre las frecuencias acumuladas observadas (AO_i) y las frecuencias acumuladas esperadas (AE_i), partiendo de la hipótesis nula de que los datos se ajustan a una distribución determinada. La fórmula del test es:

$$D = \max\left|AO_i - AE_i\right|$$

Este test fue recalculado para una distribución Normal estudiando las frecuencias esperadas a partir de la media y varianza de la muestra (Lilliefors, 1967) y se conoce como test *K-S-L*. Sin embargo, su aplicación es limitada cuando existen pocos datos (son necesarias 100 observaciones para distinguir entre una Normal con $\mu = 0$ y $\sigma^2 = 1$ de una distribución uniforme entre $-\sqrt{3}$ y $\sqrt{3}$).

Una vez calculado el estadístico *D* se contrasta con un valor *D* crítico para el nivel de significación elegido que aparece en la Tabla 5 del Apéndice I o archivo **Tabla 5.doc** (para el test *K-S*) y en la Tabla 6 del Apéndice I o archivo **Tabla 6.doc** (para el test *K-S-L*). La hipótesis nula se acepta cuando *D* observado es inferior al valor de *D* tabulado o cuando el valor *p* del contraste es mayor que el nivel de significación (por ejemplo mayor que 0,05).En los Cuadros VI.1 y VI.2 se indica cómo realizar este análisis.

VI.3. Test de Shapiro-Wilk

Es la prueba más recomendable para testar la normalidad de una muestra, sobre todo si se trabaja con un número pequeño de datos (*n* < 30). Solo se utiliza para testar la distribución Normal (Shapiro & Wilk, 1965). Se basa en medir el ajuste de los datos a una recta probabilística Normal (Figura VI.1). Si el ajuste fuera perfecto los puntos formarían una recta de 45º (frecuencia observada igual a frecuencia esperada). El estadístico de contraste se expresa por medio de la siguiente ecuación:

$$W = \frac{1}{\sum_{j=1}^{n}(x_j - \mu)^2}\left[\sum_{j=1}^{h} a_{j,n}(x_{(n-j+1)} - x_j)\right]^2$$

donde n es el número de datos, x_j es el dato en orden ascendente de muestra que ocupa el lugar j, μ es la media, h es $n/2$ si n es par o $(n\text{-}1)/2$ si n es impar y $a_{j,n}$ es un valor tabulado.

Una vez calculado el estadístico W se contrasta con un valor W crítico para el nivel de significación elegido. Como este estadístico mide el ajuste a una recta y no la distancia a la distribución Normal (como era en los casos anteriores), y se puede interpretar de forma aproximada como un coeficiente de correlación entre los valores observados y esperados (un valor próximo a 1 indica buen ajuste, y próximo a 0 malo), la hipótesis nula se acepta cuando el valor W es superior al valor de contraste tabulado (valor de ajuste muy alto). En el Cuadro VI.2 se indica el método para realizar este análisis.

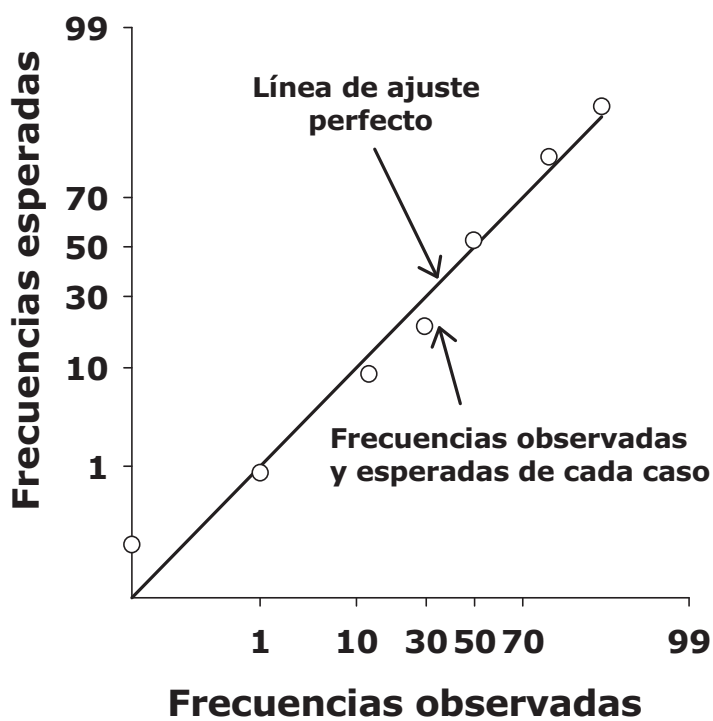


Figura VI.1. Representación de las probabilidades esperadas y observadas de una muestra junto con la línea de ajuste perfecto a una distribución normal.

CUADRO VI.1. Ajuste de distribuciones con el test χ^2 y el test de Kolmogorov-Smirnov

Aplicación con R.

EJEMPLO. En el archivo **Cuadro VI.1.V.csv** hay datos, en dos zonas diferentes, de la abundancia de dos especies de caracoles de mar herbívoras (*Littorina littorea* y *Littorina saxatilis*) y una especie carnívora (*Nucella lapillus*). Se desea conocer si la distribución es aleatoria en las distintas zonas, es decir, si siguen una distribución de Poisson.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro VI.1.R**, que se muestra en la siguiente ventana. Solo es necesario realizar, para una aplicación con otros datos diferentes, los cambios que se desee en la sección modificable por el usuario. Para que este *script* funcione correctamente es necesario tener instalado el paquete "MASS".

```
# Cuadro VI.1

# Ajustes de distribuciones con el test X^2 y el test de Kolmogorov-Smirnov

#########################################################
# Sección modificable por el usuario
#########################################################

# Lectura de la base de datos
datos<-read.csv2("Cuadro VI.1.V.csv",enc="latin1")

# Selección de las variables de interés
# varInteres<-c("Zona1")
varInteres<-c("Zona1","Zona2")

# Selección de categorización de las variables
# Si no se utilizan variables de agrupación obligatoriamente
# se exige que se coloce varAgrupación como NULL
# varAgrupacion<-NULL
varAgrupacion<-c("Especie")
#varAgrupacion<-c("Género","Especie")


# Distribución para la cual se quiere hacer el hacer el ajuste
# 1. Binomial
# 2. Poisson
distribucion<-2

# ¿Combine categorías inferior a 5?
combinar<-"si"

# Gráfica comparativa
grafica<-"si"

# Nombre del archivo de salida
ArchivodeSalida<-"Salida Cuadro VI.1.V.csv"

# Titulo genérico del eje x.
tituloEjeX<-"Abundancia"
```

CUADRO VI.1. *(Continuación)*[R]

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro VI.1.V.csv**.

```
datos<-read.csv2("Cuadro VI.1.V.csv",enc="latin1")
```

Paso 2. Se seleccionan las variables cuya distribución queremos comprobar. En el ejemplo se desea ver si la distribución de los caracoles es al azar en las zonas 1 y 2.

```
# varInteres<-c("Zona1")
varInteres<-c("Zona1","Zona2")
```

Paso 3. Es posible hacer los cálculos estadísticos agrupando las variables de la forma que se quiera. En el caso de seleccionar «NULL», no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna. En el ejemplo se selecciona que los estadísticos se hagan a nivel de especie.

```
# varAgrupacion<-NULL
varAgrupacion<-c("Especie")
#varAgrupacion<-c("Género","Especie")
```

Paso 4. Es posible hacer el contraste para ver si los datos tienen una distribución Binomial o de Poisson. En el ejemplo nos interesa ver la distribución de Poisson.

```
# 1. Binomial
# 2. Poisson
distribucion<-2
```

Paso 5. Es aconsejable que si el número de casos esperados dentro de una clase es menor que 5 se combinen varias clases en una, hasta conseguir que todas tengan una frecuencia esperada mayor o igual a 5. Por eso es conveniente marcar con «si» a la pregunta de combinar las categorías si el valor es menor que 5, así se puede tener información tanto de la prueba sin combinar (contraste χ^2 no ajustado) como combinando los datos (contraste χ^2 ajustado).

```
combinar<-"si"
```

Paso 6. Es posible representar las gráficas de las distribuciones esperada y real si marcamos «si» a la pregunta de si queremos la gráfica comparativa.

```
grafica<-"si"
```

CUADRO VI.1. *(Continuación)*[R]

Paso 7. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados y la leyenda de eje X de las gráficas.

```
ArchivodeSalida<-"Salida Cuadro VI.1.V.csv"
tituloEjeX<-"Abundancia"
```

Paso 8. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*. Los datos se muestran a continuación.

nombres	Especie	variable	D de K-S	Valor-p K-S	X Chisq2	Grados de libertad ajustado	Valor-p Chisq2 ajustado
L. littorea.Zona1	L. littorea	Zona1	0,28872113	0,00214984	12,9149482	4	0,011698936
L. littorea.Zona2	L. littorea	Zona2	0,15153169	0,45445897	5,12697276	3	0,162732211
L. saxatilis.Zona1	L. saxatilis	Zona1	0,29565653	0,00743713	12,0614729	3	0,00717552
L. saxatilis.Zona2	L. saxatilis	Zona2	0,27091062	0,19084638	NA	NA	NA
N. lapillus.Zona1	N. lapillus	Zona1	0,31581487	1,27E-05	15,7339841	5	0,007646062
N. lapillus.Zona2	N. lapillus	Zona2	0,2213481	0,01812666	28,0103039	4	1,24E-05

En el caso del test χ^2, la prueba ajustada agrupa las categorías para conseguir frecuencias esperadas no menores que cinco. Cuando los resultados de las pruebas no ajustada y ajustada son distintos, eso significa que ha sido necesario agrupar y, por lo tanto, que la prueba no ajustada no es válida. Los resultados muestran que en todos los casos los valores de probabilidad para el test χ^2 de la prueba ajustada son diferentes de la no ajustada y, por tanto, en todos los casos se tiene que considerar la prueba ajustada como válida.

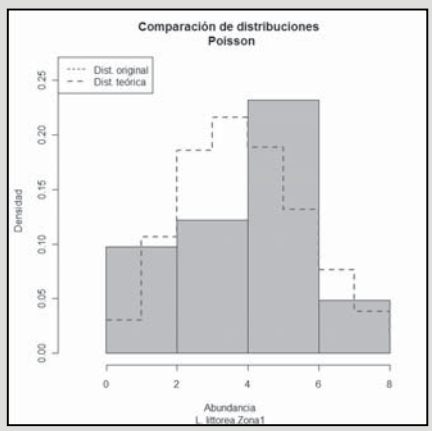
Para *Littorina littorea* en la zona 1, el valor *p* del contraste χ^2 es menor que 0,05 en la prueba ajustada (p = 0,011), por tanto se rechaza la hipótesis de que la distribución de esta especie es aleatoria en la zona 1, es decir, no sigue una distribución de Poisson. El test de Kolmogorov-Smirnov (K-S) corrobora lo antes mencionado ya que *p* es 0,002 (menor que 0,05). Por el contrario, para la misma especie en la zona 2, el valor *p* del contraste χ^2 es mayor que 0,05 en la prueba ajustada (p = 0,162), lo cual indica que la distribución es al azar en esa zona. El test K-S corrobora lo antes mencionado ya que *p* es 0,454 (mayor que 0,05).

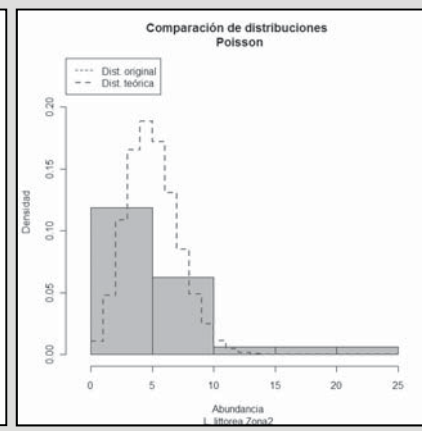
En el caso de *Littorina saxatilis*, el valor *p* del contraste χ^2 es menor que 0,05 en la zona 1 (p = 0,007 ajustado), indicando que no sigue una distribución de Poisson, y no se puede estimar en la zona 2. Con el test de K-S la probabilidad también es menor que 0,05 en la zona 1 (p = 0,007), indicando que la distribución no es de Poisson. El test K-S es más apropiado que el test χ^2 para variables cuantitativas, por lo que daremos validez al test K-S. En cualquier caso, si decide aplicar varias pruebas válidas con el mismo fin, como en este ejemplo que usamos los test χ^2 y K-S, y alguna de ellas rechaza la hipótesis nula y otras no, se debe rechazar la hipótesis nula.

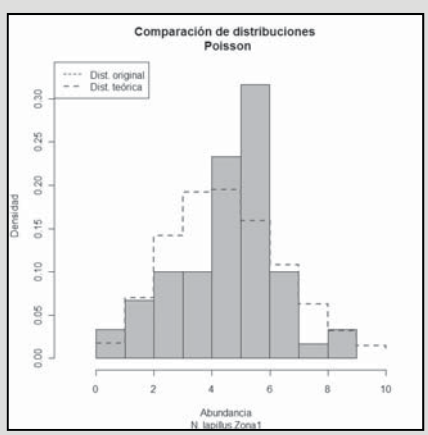
Para *Nucella lapillus*, en las dos zonas, y tanto el χ^2 de las pruebas ajustadas como el test K-S, el valor *p* es menor que 0,05, rechazando la hipótesis de distribución aleatoria de Poisson.

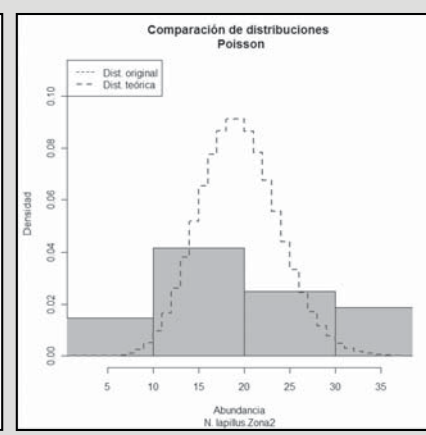
CUADRO VI.1. *(Continuación)*[R]

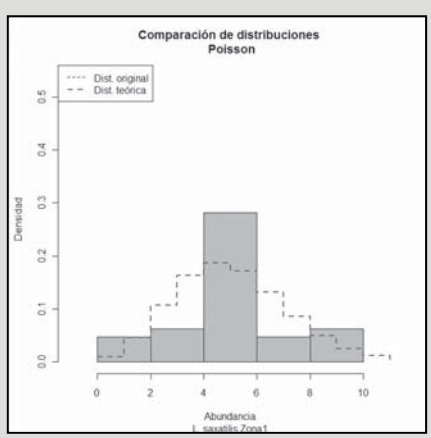
Paso 9. Para poder ver todas las gráficas, solo hay que arrastrar cada una de las ventanas, y así poder ver la gráfica que hay debajo.

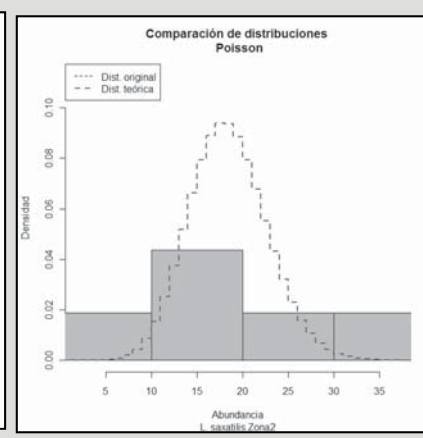










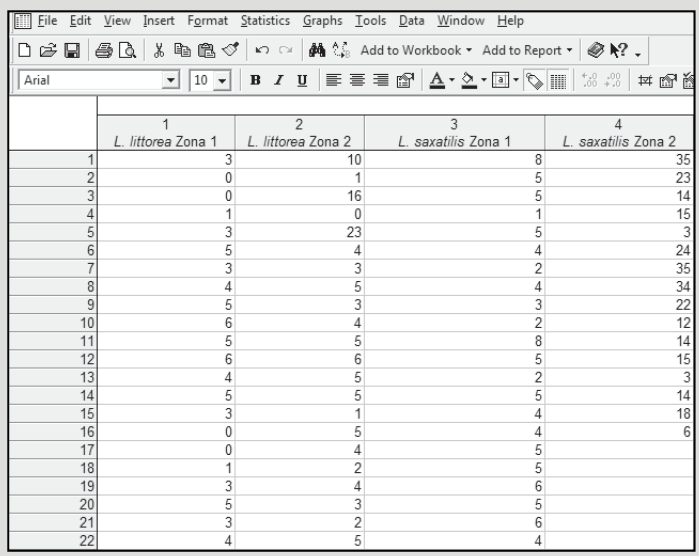


CUADRO VI.1. *(Continuación)*^{STATISTICA}

Aplicación con STATISTICA.

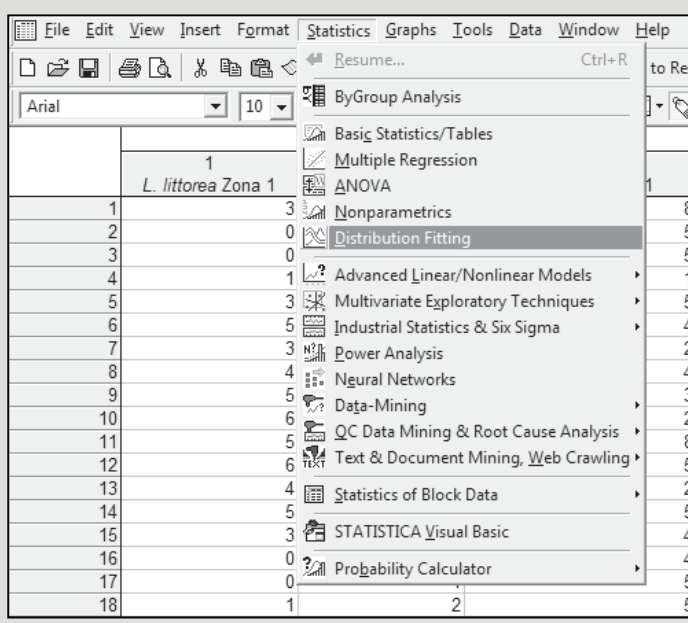
EJEMPLO. Vamos a utilizar los mismos datos que en el caso explicado con R, los cuales se encuentran en el archivo **Cuadro VI.1.sta**, pero vamos a ver el caso de la especie *Nucella lapillus* en la zona 2.

Paso 1. En STATISTICA los datos hay que organizarlos en columnas como se muestra en la siguiente ventana.



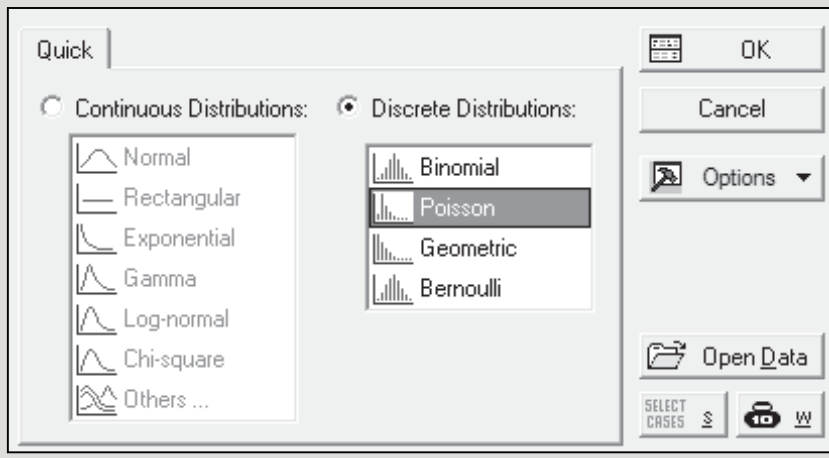
	1 *L. littorea* Zona 1	2 *L. littorea* Zona 2	3 *L. saxatilis* Zona 1	4 *L. saxatilis* Zona 2
1	3	10	8	35
2	0	1	5	23
3	0	16	5	14
4	1	0	1	15
5	3	23	5	3
6	5	4	4	24
7	3	3	2	35
8	4	5	4	34
9	5	3	3	22
10	6	4	2	12
11	5	5	8	14
12	6	6	5	15
13	4	5	2	3
14	5	5	5	14
15	3	1	4	18
16	0	5	4	6
17	0	4	5	
18	1	2	5	
19	3	4	6	
20	5	3	5	
21	3	2	6	
22	4	5	4	

Paso 2. Se selecciona «Statistics» ⇨ «Distribution Fitting».

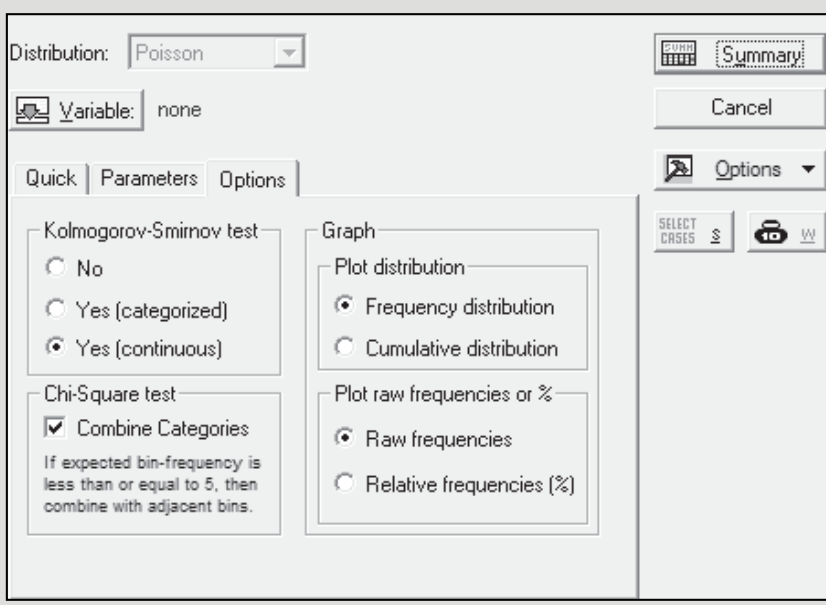


CUADRO VI.1. *(Continuación)*[STATISTICA]

Paso 3. Nos aparece la siguiente ventana donde tenemos que elegir el tipo de distribución con la que se quiere contrastar nuestros datos.



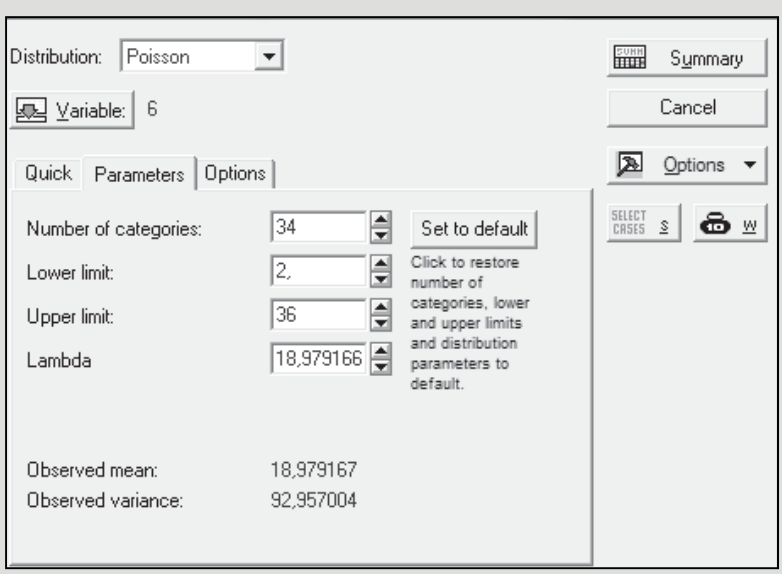
Paso 4. Pulsando «OK» en la ventana anterior aparece la siguiente ventana en la que se distingue una parte común y tres pestañas diferentes. En la pestaña «Options» se seleccionan en la parte izquierda los contrastes que queremos efectuar. En primer lugar, debemos indicar si queremos realizar un test Kolmogorov-Smirnov, y en ese caso se debe seleccionar la opción «Yes (continuous)». También se debe marcar "Combine Categories", para que al realizar el test χ^2 agrupe las categorías con frecuencias inferiores a 5.



CUADRO VI.1. *(Continuación)*[STATISTICA]

Paso 5. En el bloque común de la ventana anterior, pulsando sobre la pestaña «Variable» aparece la siguiente ventana donde se especifica la variable que se quiere analizar.

Paso 6. En la pestaña «Parameters» hay que especificar el «Lower limit» que es el valor mínimo de la serie de datos menos 1 (en este ejemplo es 2), salvo que el valor sea 0 que se dejaría entonces como 0, el «Upper limit» que es el valor máximo de la serie de datos más 1 (en este ejemplo es 36), y «Number of categories» que es la diferencia entre el «Lower limit» y el «Upper limit» (en este ejemplo es 34).



CUADRO VI.1. *(Continuación)*[STATISTICA]

Paso 7. Al pulsar «Summary» aparecerá la siguiente tabla de resultados.

Variable: Nucella lapillus Zona 2, Distribution: Poisson, Lambda = 18,9 (Cuadro VI.1.sta)
Kolmogorov-Smirnov d = 0,22135, p < 0,05
Chi-Square = 28,01031, df = 4 (adjusted) , p = 0,00001

Category	Observed Frequency	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
<= 2,00000	0	0	0,00000	0,0000	0,000055	0,00005	0,000114	0,0001	-0,00005
3,00000	4	4	8,33333	8,3333	0,000313	0,00037	0,000652	0,0008	3,99969
4,00000	0	4	0,00000	8,3333	0,001485	0,00185	0,003093	0,0039	-0,00148
5,00000	0	4	0,00000	8,3333	0,005635	0,00749	0,011740	0,0156	-0,00564
6,00000	3	7	6,25000	14,5833	0,017825	0,02531	0,037135	0,0527	2,98218
7,00000	0	7	0,00000	14,5833	0,048329	0,07364	0,100684	0,1534	-0,04833
8,00000	0	7	0,00000	14,5833	0,114654	0,18830	0,238863	0,3923	-0,11465
9,00000	0	7	0,00000	14,5833	0,241783	0,43008	0,503714	0,8960	-0,24178
10,00000	1	8	2,08333	16,6667	0,458883	0,88896	0,956007	1,8520	0,54112
11,00000	0	8	0,00000	16,6667	0,791748	1,68071	1,649474	3,5015	-0,79175
12,00000	3	11	6,25000	22,9167	1,252226	2,93293	2,608803	6,1103	1,74777
13,00000	0	11	0,00000	22,9167	1,828169	4,76110	3,808686	9,9190	-1,82817
14,00000	6	17	12,50000	35,4167	2,478366	7,23947	5,163263	15,0822	3,52163
15,00000	4	21	8,33333	43,7500	3,135822	10,37529	6,532963	21,6152	0,86418
16,00000	0	21	0,00000	43,7500	3,719704	14,09500	7,749383	29,3646	-3,71970
17,00000	3	24	6,25000	50,0000	4,152773	18,24777	8,651610	38,0162	-1,15277
18,00000	3	27	6,25000	56,2500	4,378662	22,62643	9,122212	47,1384	-1,37866
19,00000	0	27	0,00000	56,2500	4,373860	27,00029	9,112209	56,2506	-4,37386
20,00000	0	27	0,00000	56,2500	4,150604	31,15089	8,647091	64,8977	-4,15060
21,00000	2	29	4,16667	60,4167	3,751197	34,90209	7,814994	72,7127	-1,75120
22,00000	3	32	6,25000	66,6667	3,236118	38,13821	6,741913	79,4546	-0,23612
23,00000	1	33	2,08333	68,7500	2,670381	40,80859	5,563293	85,0179	-1,67038
24,00000	6	39	12,50000	81,2500	2,111736	42,92033	4,399449	89,4173	3,88826
25,00000	0	39	0,00000	81,2500	1,603159	44,52348	3,339915	92,7573	-1,60316
26,00000	0	39	0,00000	81,2500	1,170254	45,69374	2,438028	95,1953	-1,17025
27,00000	0	39	0,00000	81,2500	0,822610	46,51635	1,713770	96,9091	-0,82261
28,00000	0	39	0,00000	81,2500	0,557587	47,07394	1,161639	98,0707	-0,55759
29,00000	0	39	0,00000	81,2500	0,364915	47,43885	0,760240	98,8309	-0,36492
30,00000	0	39	0,00000	81,2500	0,230859	47,66971	0,480957	99,3119	-0,23086
31,00000	0	39	0,00000	81,2500	0,141339	47,81105	0,294457	99,6064	-0,14134
32,00000	0	39	0,00000	81,2500	0,083828	47,89488	0,174642	99,7810	-0,08383
33,00000	0	39	0,00000	81,2500	0,048212	47,94309	0,100441	99,8814	-0,04821
34,00000	4	43	8,33333	89,5833	0,026912	47,97000	0,056067	99,9375	3,97309
< Infinity	5	48	10,41667	100,0000	0,029998	48,00000	0,062496	100,0000	4,97000

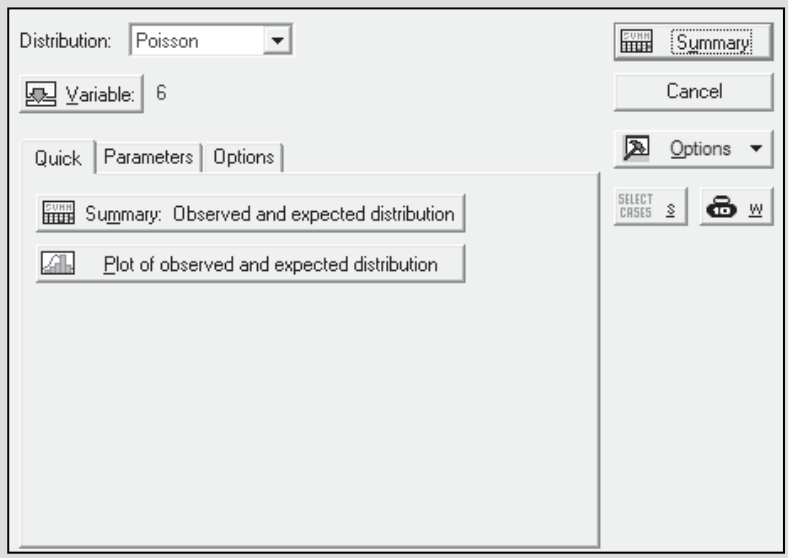
En la parte superior aparecen el valor del test χ^2 y el del test Kolmogorov-Smirnov (K-S).

STATISTICA no da el valor de *p* en el caso del test K-S, ya que no lo calcula, sino que compara el valor *d* (en este caso 0,221) con una tabla. Simplemente nos dice si *p* es < 0,01 o < 0,05, en cuyo caso debemos rechazar la hipótesis nula, y la distribución no es la del contraste (en este ejemplo de Poisson). Si es mayor que 0,05 no pone nada y significa que podemos aceptar que la distribución es la del contraste.

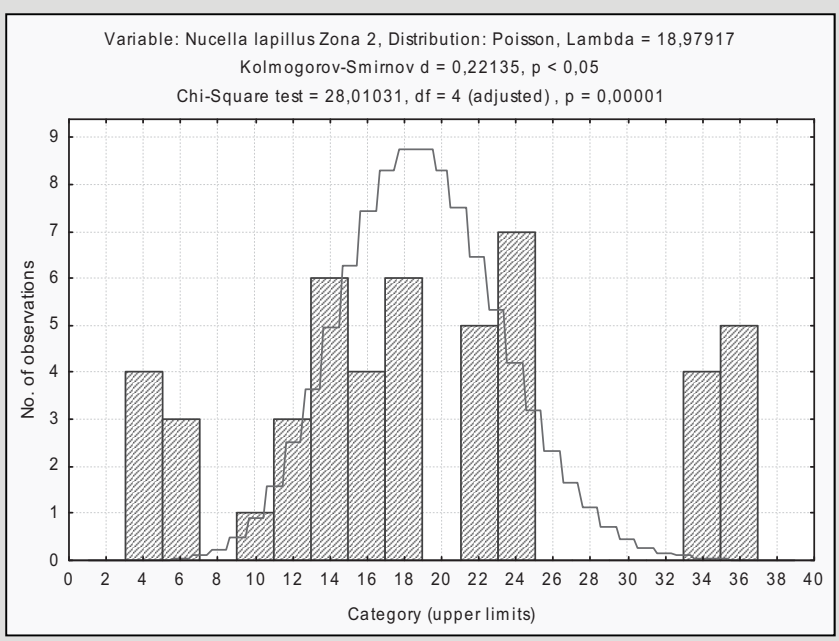
Los resultados obtenidos con el test χ^2 corroboran lo obtenido con el test K-S. Como podemos observar para nuestros datos el valor χ^2 de la muestra es 28,01 con 4 grados de libertad. Como *p* es menor que 0,05 ($p < 0,001$), se rechaza la hipótesis de que la abundancia se ajusta a una distribución aleatoria de Poisson.

Paso 8. Al pulsar la pestaña «Quick» aparece la siguiente ventana. La tabla de datos anterior se puede obtener pulsando «Summary: Observed and expected distribution».

CUADRO VI.1. *(Continuación)* STATISTICA



En «Plot of observed and expected distribution», se obtiene el gráfico que se muestra a continuación. En la parte superior del gráfico aparecen, al igual que en la tabla anterior, los estadísticos de contraste y en el gráfico se muestran las frecuencias observadas como barras verticales y las frecuencias esperadas como una línea, pudiéndose comprobar las diferencias entre los valores observados y esperados que nos han llevado a rechazar la hipótesis nula.

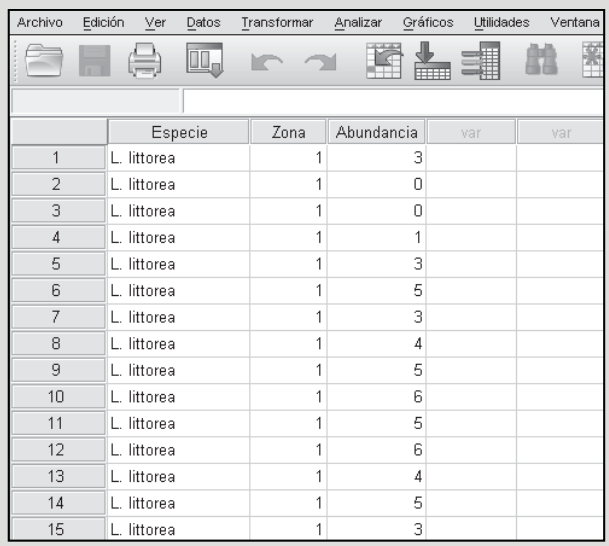

Variable: Nucella lapillus Zona 2, Distribution: Poisson, Lambda = 18,97917
Kolmogorov-Smirnov d = 0,22135, p < 0,05
Chi-Square test = 28,01031, df = 4 (adjusted) , p = 0,00001
No. of observations
0
1
2
3
4
5
6
7
8
9
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Category (upper limits)

CUADRO VI.1. *(Continuación)*[SPSS]

Aplicación con SPSS.

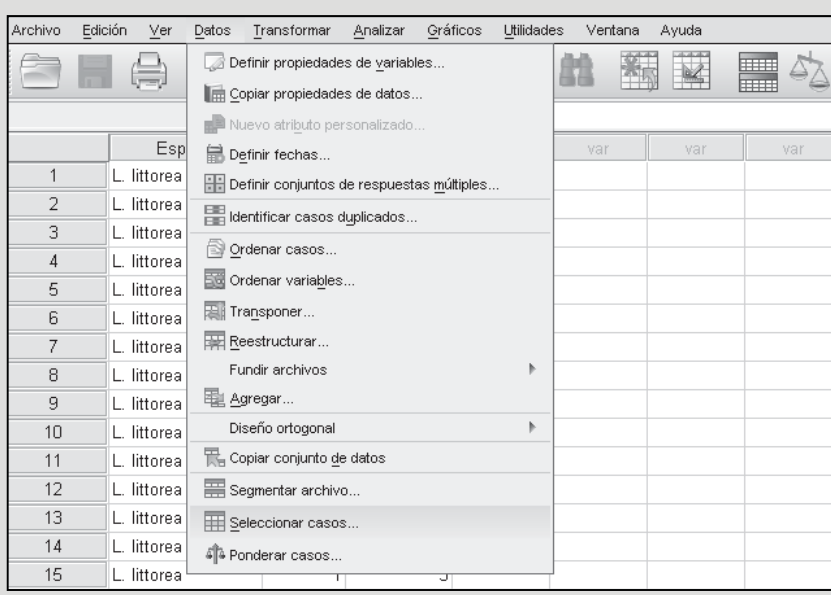
EJEMPLO. Vamos a utilizar el mismo ejemplo de antes, el cual se encuentra en el archivo **Cuadro VI.1.sav**, pero vamos a ver la distribución de la especie *Littorina littorea* en la zona 2.

Paso 1. En SPSS los datos hay que organizarlos como se muestra en el archivo y en la siguiente ventana.



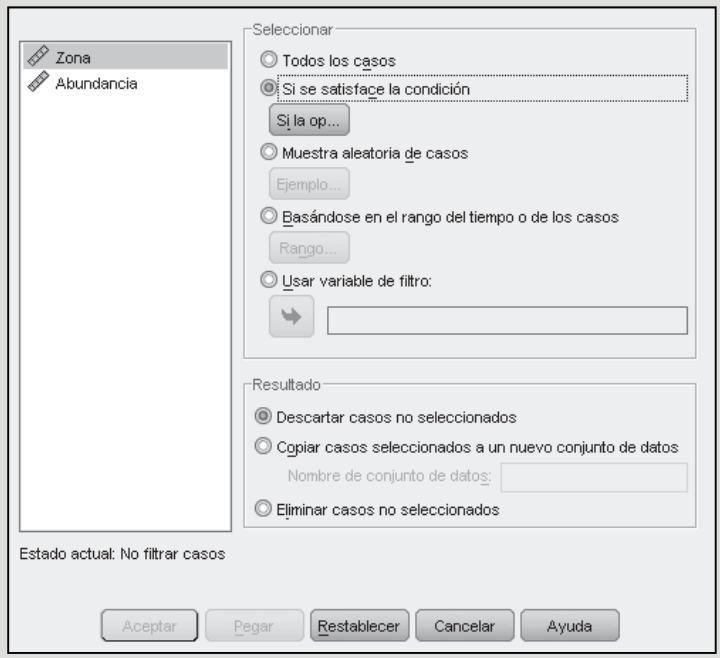
	Especie	Zona	Abundancia	var	var
1	L. littorea	1	3		
2	L. littorea	1	0		
3	L. littorea	1	0		
4	L. littorea	1	1		
5	L. littorea	1	3		
6	L. littorea	1	5		
7	L. littorea	1	3		
8	L. littorea	1	4		
9	L. littorea	1	5		
10	L. littorea	1	6		
11	L. littorea	1	5		
12	L. littorea	1	6		
13	L. littorea	1	4		
14	L. littorea	1	5		
15	L. littorea	1	3		

Paso 2. En primer lugar hay que seleccionar el grupo que vamos a analizar. Para ello se elige en el menú «Datos» ⇨ «Seleccionar casos…».

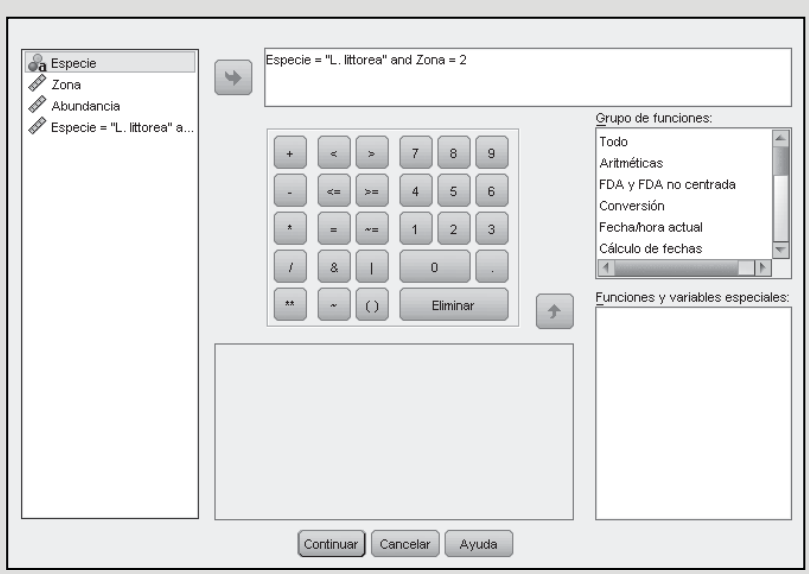


CUADRO VI.1. *(Continuación)*[SPSS]

Paso 3. En la ventana que aparece se marca «Si satisface la condición» y se pulsa el botón «Si la op…».

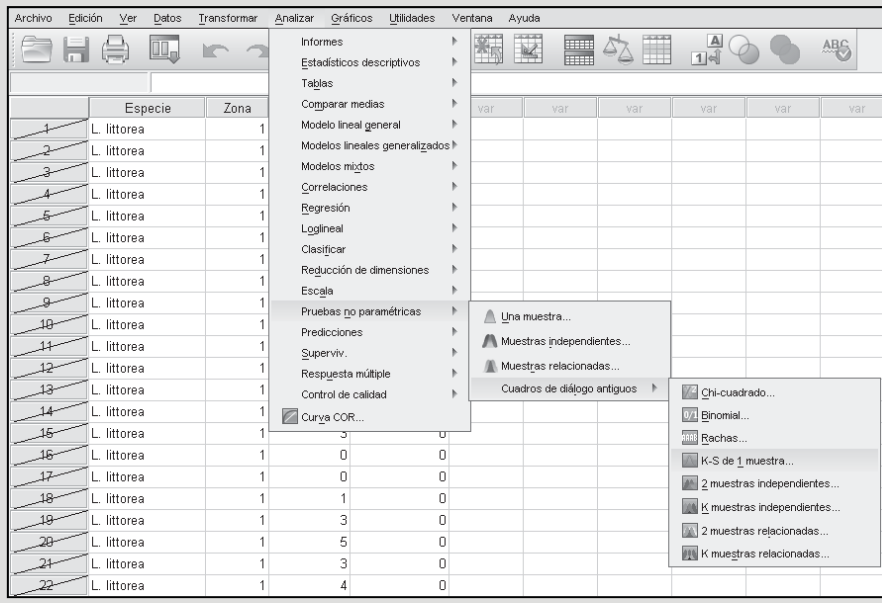


Paso 4. A continuación se seleccionan los grupos con los que se va a trabajar y, por ello, escribimos «Especie="L. littorea" and Zona=2», de tal forma que determinaremos si la distribución de la especie *L. littorea* en la zona 2 es aleatoria, es decir, tiene una distribución de Poisson.

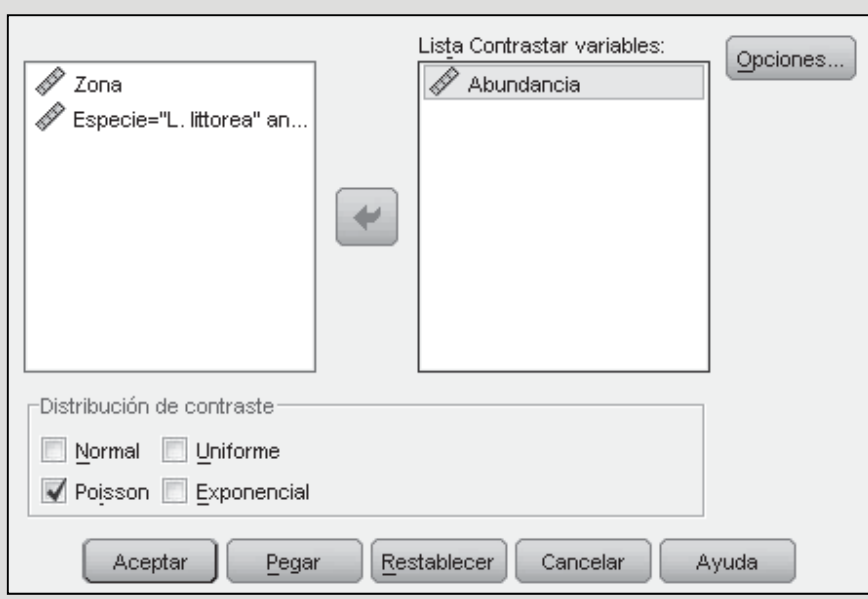


CUADRO VI.1. *(Continuación)*[SPSS]

Paso 5. Con SPSS la prueba se hace de forma más fácil usando el test de Kolmogorov-Smirnov. Para ello se selecciona «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «K-S de 1 muestra…».

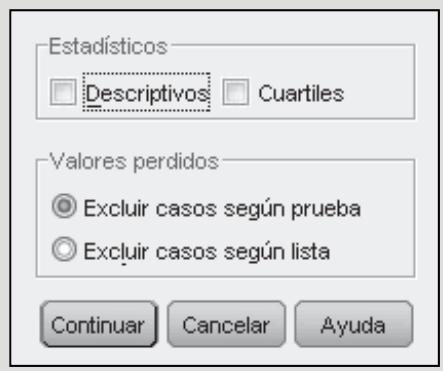


Paso 6. En la siguiente ventana marcamos «Poisson» y en «Lista Contrastar variables:» se especifican las variables que se quieren contrastar con la distribución de Poisson.



CUADRO VI.1. *(Continuación)*[SPSS]

En la pestaña «Opciones» nos aparece la siguiente ventana donde podemos elegir la presentación de algunos resultados adicionales (estadísticos descriptivos, cuartiles) y el tratamiento de los valores perdidos. El método «Excluir casos según lista» excluye todos los casos con algún valor perdido o faltante (*missing*), mientras que el método «Excluir casos según prueba» realiza los cálculos de cada variable utilizando todos los casos en los que exista dato (aunque falte para otra variable). En general debemos elegir este último procedimiento.



Paso 7. Los resultados se muestran en la siguiente ventana. El estadístico "d" de Kolmogorov es la diferencia más extrema absoluta, es decir 0,152. SPSS, a diferencia de STATISTICA, no compara simplemente este valor con los existentes en la tabla de K-S para decidir si *p* es o no menor que el nivel de significación, sino que calcula un valor *p* aproximado, en este ejemplo 0,454, que por ser mayor que 0,05 nos permite aceptar que los datos tienen una distribución de Poisson. Es necesario mencionar que SPSS no da la opción de elegir como STATISTICA y al realizar el test a veces los datos los considera como categóricos y otras veces como continuos. Esto significa que realizar la prueba de bondad de ajuste de Kolmogorov-Smirnov con SPSS puede dar lugar a obtener conclusiones ligeramente diferentes que si se hace con STATISTICA, debido a la distinta forma de agrupar los valores de la variable.

Prueba de Kolmogorov-Smirnov para una muestra

		Abundancia
N		32
Parámetro de Poisson[a,,b]	Media	4,56
Diferencias más extremas	Absoluta	,152
	Positiva	,152
	Negativa	-,075
Z de Kolmogorov-Smirnov		,857
Sig. asintót. (bilateral)		,454

a. La distribución de contraste es la de Poisson.

b. Se han calculado a partir de los datos.

CUADRO VI.2. Contraste de Normalidad con los test de Kolmogorov-Smirnov y de Shapiro-Wilk

Aplicación con R.

EJEMPLO. Vamos a utilizar el mismo ejemplo del archivo **Cuadro VI.1.V.csv** que son datos de abundancias de tres especies de caracoles de mar, *Littorina littorea, Littorina saxatilis* y *Nucella lapillus*. Se desea conocer si la distribución de las especies es Normal en cada una de las dos zonas estudiadas.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el *script* **Cuadro VI.2.R**, que se muestra en la siguiente ventana. Lo único que hay que hacer es, para una aplicación con otros datos, cambiar la sección modificable por el usuario.

Para que este *script* funcione correctamente es necesario tener instalados los paquetes "nortest", "MASS" y "e1071".

```
# Cuadro VI.2
# Contraste de Normalidad con
# el test de Kolmogorov-Smirnov
# y el test de Shapiro-Wilk
#########################################################
# Sección modificable por el usuario
#########################################################

datos<-read.csv2("Cuadro VI.1.V.csv",enc="latin1")

# Selección de las variables de interés
# varInteres<-c("Zona1")
 varInteres<-c("Zona1","Zona2")

# Selección de categorización de las variables
# Si no se utilizan variables de agrupación obligatoriamente
# se exige que se coloce varAgrupación como NULL
# varAgrupacion<-NULL
varAgrupacion<-c("Especie")
# varAgrupacion<-c("Género","Especie")

# ¿Se dibuja histograma?
graficaHist<-"si"

# Número de intervalos, para la gráfica.
ni<-12

# Nombre del archivo de salida
ArchivodeSalida<-"Salida Cuadro VI.2.V.csv"
```

CUADRO VI.2. *(Continuación)*[R]

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro VI.1.V.csv**.

```
datos<-read.csv2("Cuadro VI.1.V.csv",enc="latin1")
```

Paso 2. Se seleccionan las variables a las que se quiere aplicar las pruebas. En el ejemplo se desea ver si la distribución de los caracoles es Normal en las zonas 1 y 2.

```
# varInteres<-c("Zona1")
varInteres<-c("Zona1","Zona2")
```

Paso 3. Es posible hacer los cálculos estadísticos agrupando las variables de la forma que se quiera. En el caso de seleccionar «NULL», no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna. En el ejemplo se selecciona que las pruebas se hagan a nivel de especie.

```
# varAgrupacion<-NULL
varAgrupacion<-c("Especie")
#varAgrupacion<-c("Género","Especie")
```

Paso 4. Es posible representar los resultados si marcamos «si» a la pregunta de si se dibuja el histograma y también es posible definir el número de intervalos de la gráfica.

```
graficaHist<-"si"
ni<-12
```

Paso 5. Por último, es necesario especificar el nombre del archivo donde se guardarán todos los cálculos realizados.

```
ArchivodeSalida<-"Salida Cuadro VI.2.V.csv"
```

Paso 6. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*. Los resultados se muestran a continuación.

	nombres	Especie	variable	n	Asimetría	D de K-S	Valor-p K-S	D de K-S Lilliefors	Valor-p K-S Lilliefors
2	L. littorea.Zona1	L. littorea	Zona1	41	-0,52822325	0,18395582	0,12468631	0,183955816	0,001266339
3	L. littorea.Zona2	L. littorea	Zona2	32	2,50715703	0,30527239	0,005138	0,30527239	3,28E-08
4	L. saxatilis.Zona1	L. saxatilis	Zona1	32	-0,07679185	0,20483028	0,13637883	0,204830284	0,001486642
5	L. saxatilis.Zona2	L. saxatilis	Zona2	16	0,32062788	0,17422108	0,71641553	0,174221083	0,21716141
6	N. lapillus.Zona1	N. lapillus	Zona1	60	-0,32175659	0,18484964	0,03313438	0,184849635	2,43E-05
7	N. lapillus.Zona2	N. lapillus	Zona2	48	0,21846098	0,12787722	0,41240806	0,127877217	0,047854752

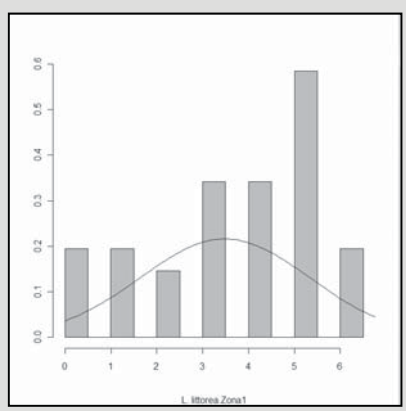
A excepción de la especie *L. saxatilis* en la zona 2 que, tanto el test K-S (p = 0,217) como Shapiro-Wilk (p = 0,164) muestran una probabilidad mayor que 0,05 y, por tanto, la abundancia de esta especie en la zona 2 tiene una distribución Normal, en todos los otros casos las probabilidades de los test K-S o Shapiro-Wilk son menores que 0,05, por tanto, se rechaza la hipótesis nula de que los datos de abundancia sigan una distribución Normal.

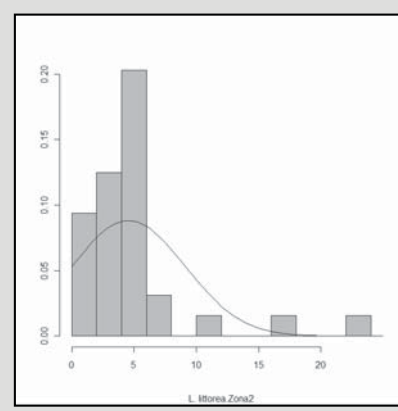
CUADRO VI.2. *(Continuación)*[R]

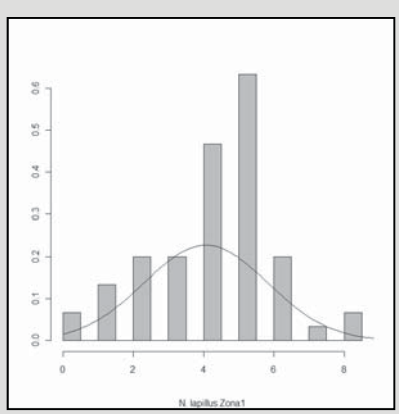
Los resultados también muestran la asimetría de cada una de las series de datos ya que, en el supuesto de que la población no tenga una distribución Normal, es posible transformar los datos como se explicará en el Capítulo VII, y la asimetría nos indica el tipo de transformación que hay que realizar.

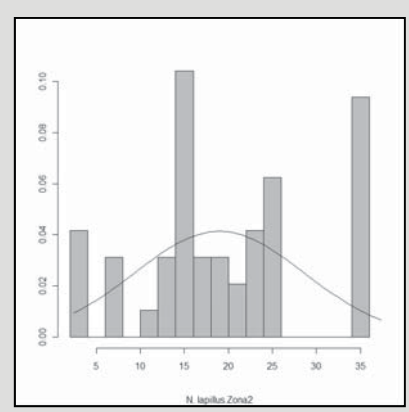
Solamente existe una discrepancia al comparar el test K-S con el test Shapiro-Wilk. La especie *L. saxatilis* en la zona 1 no tiene una distribución Normal según el test K-S (p = 0,001), mientras que el test Shapiro-Wilk si muestra distribución Normal (p = 0,065). Como ya se ha mencionado en otras ocasiones, si existe discrepancia entre dos pruebas, se debe tomar una decisión conservadora, es decir, rechazar la hipótesis nula y, por tanto, en este ejemplo se rechaza que la distribución sea Normal.

Paso 7. A continuación se muestran algunas de las gráficas que salen en el análisis, que representan las frecuencias observadas como barras verticales y las frecuencias esperadas como una línea continua. Para poder ver todas las gráficas, solamente hay que arrastrar cada una de las ventanas, y así poder ver la gráfica que hay debajo.







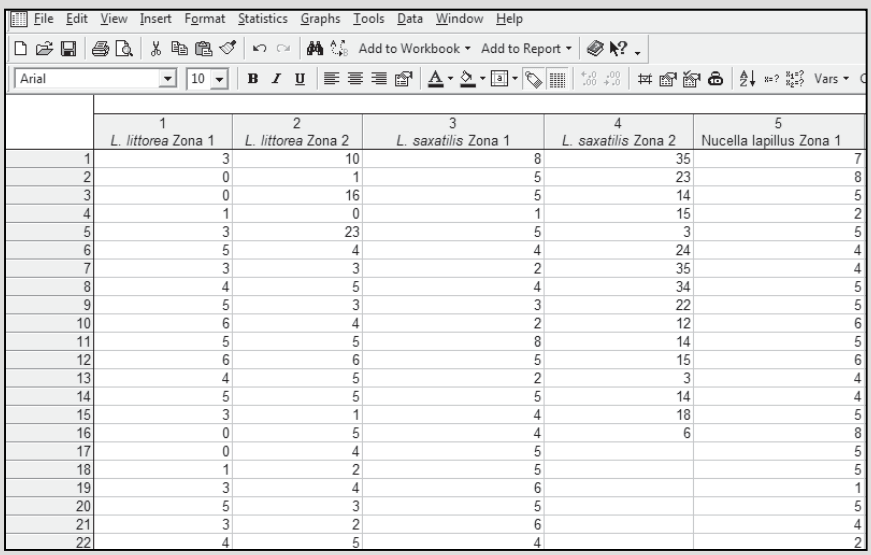


CUADRO VI.2. *(Continuación)* STATISTICA

Aplicación con STATISTICA

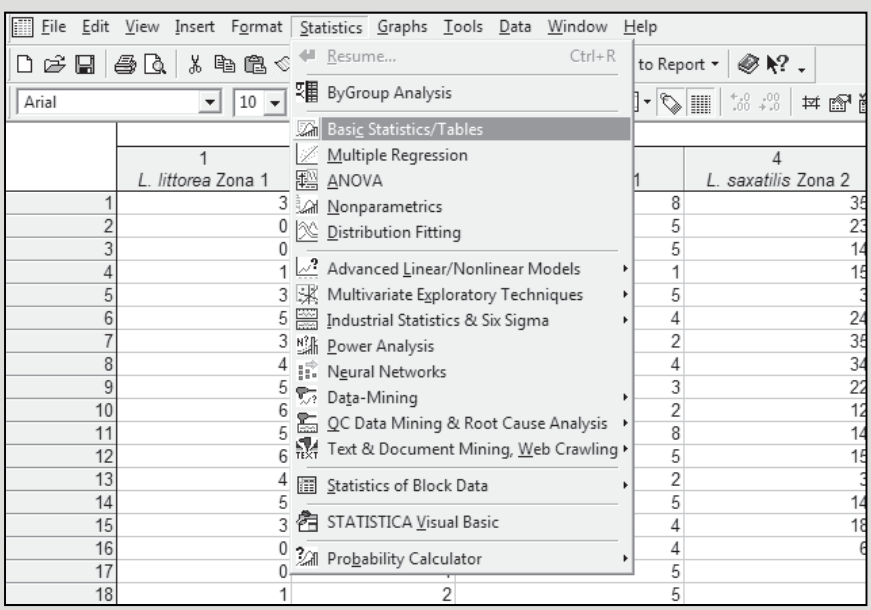
EJEMPLO. Vamos a utilizar el mismo ejemplo de antes, el cual se encuentra en el archivo **Cuadro VI.1.sta**. Vamos a determinar si la distribución de la especie *Littorina saxatilis* en la zona 1 es Normal.

Paso 1. En STATISTICA los datos hay que organizarlos en columnas como se muestra en la siguiente ventana.



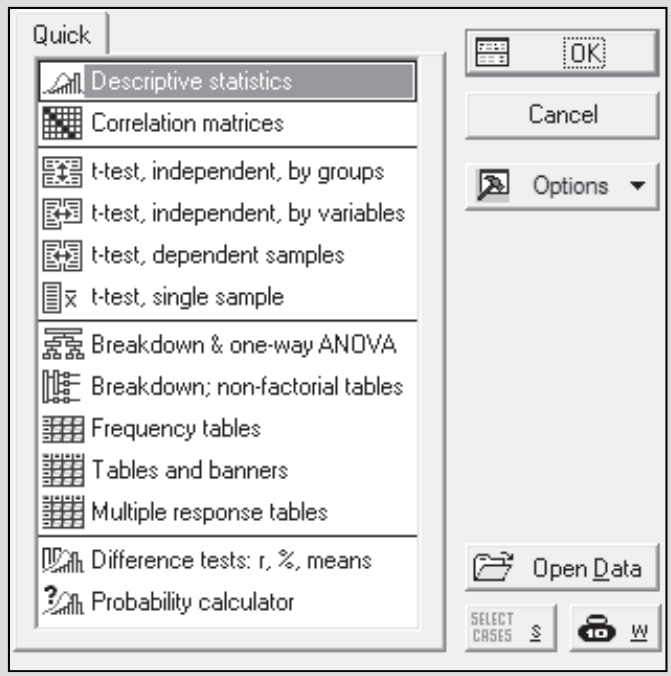
	1 *L. littorea* Zona 1	2 *L. littorea* Zona 2	3 *L. saxatilis* Zona 1	4 *L. saxatilis* Zona 2	5 Nucella lapillus Zona 1
1	3	10	8	35	7
2	0	1	5	23	8
3	0	16	5	14	5
4	1	0	1	15	2
5	3	23	5	3	5
6	5	4	4	24	4
7	3	3	2	35	4
8	4	5	4	34	5
9	5	3	3	22	5
10	6	4	2	12	6
11	5	5	8	14	5
12	6	6	5	15	6
13	4	5	2	3	4
14	5	5	5	14	4
15	3	1	4	18	5
16	0	5	4	6	8
17	0	4	5		5
18	1	2	5		5
19	3	4	6		1
20	5	3	5		5
21	3	2	6		4
22	4	5	4		2

Paso 2. Se selecciona «Statistics» ⇨ «Basic Statistics/Tables».

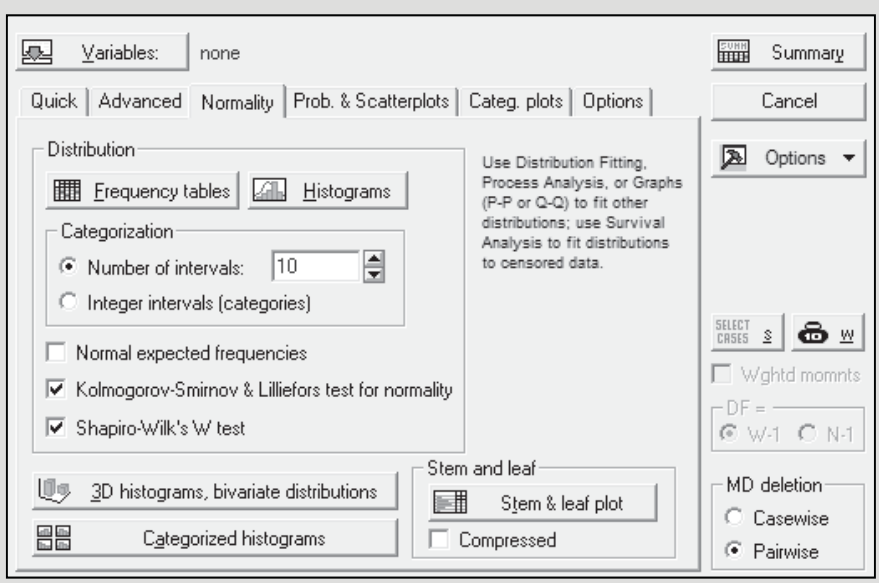


CUADRO VI.2. *(Continuación)*[STATISTICA]

Paso 3. En la ventana que aparece se selecciona «Descriptive statistics».

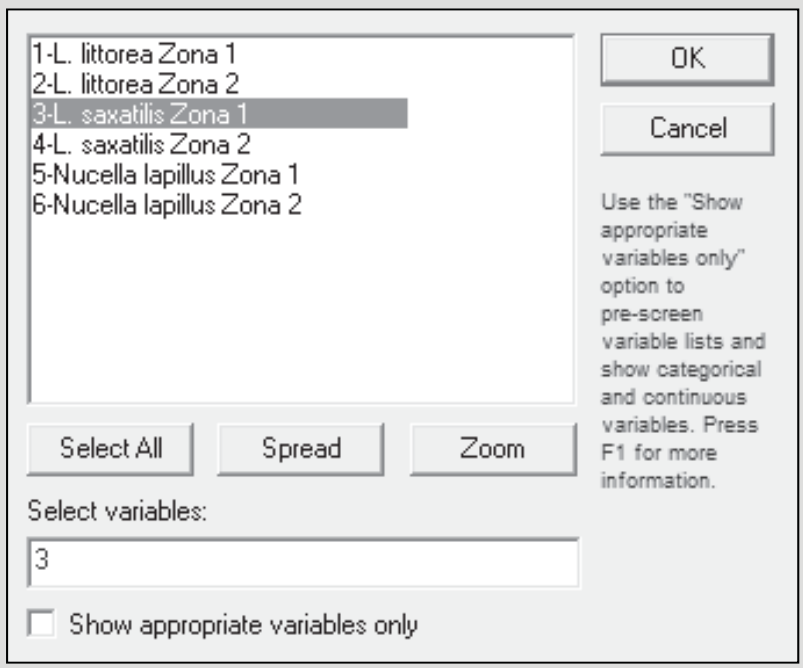


Paso 4. En la siguiente ventana, en la pestaña «Normality», se seleccionan los tests «Kolmogorov-Smirnov & Lilliefors test for normality» y «Shapiro-Wilk´s W test», y pulsamos en el botón «Variables»

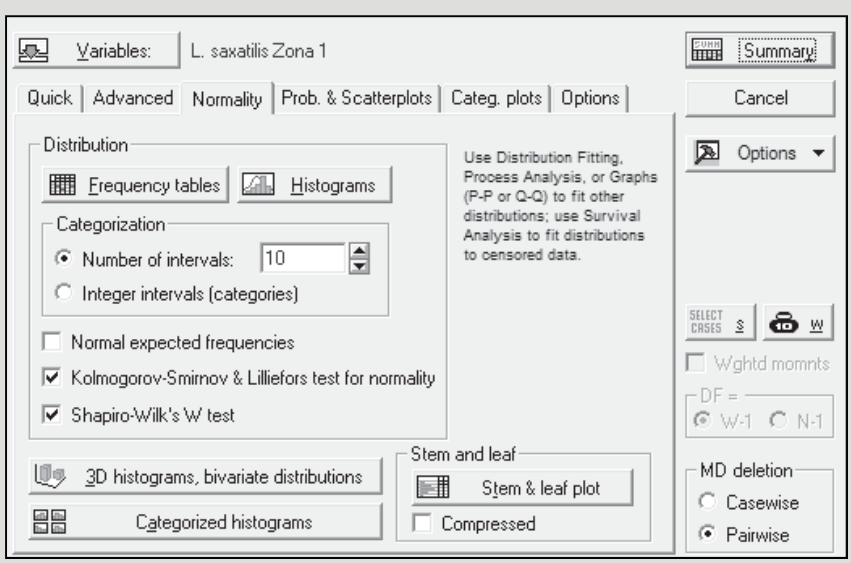


CUADRO VI.2. *(Continuación)*[STATISTICA]

Paso 5. En la siguiente ventana seleccionamos la variable para la cual se quiere averiguar si tiene una distribución Normal, en este caso la especie *L. saxatilis* en la zona 1.

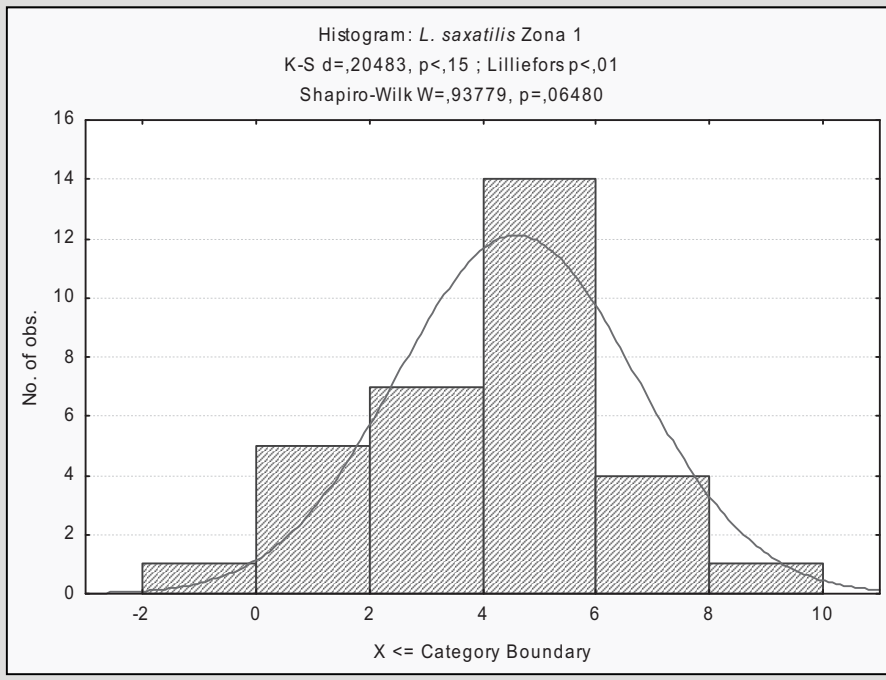


Paso 6. Por último, pulsamos el botón «Histograms».



CUADRO VI.2. *(Continuación)*[STATISTICA]

Paso 7. Los resultados se muestran en la siguiente figura.



La especie *L. saxatilis* en la zona 1 no tiene una distribución Normal según el test K-S ($p < 0,01$), aunque el dato no es exacto y no se puede saber exactamente ya que pone solo < 0,01 y, por el contrario, el test Shapiro-Wilk si muestra distribución Normal, ya que el valor de p es mayor de 0,05 ($p = 0,065$).

En este tipo de situaciones donde existe discrepancia entre las pruebas, y hay dudas, es conveniente adoptar una postura conservadora y rechazar la hipótesis nula, es decir, se rechaza la hipótesis de que los datos se ajustan a una distribución Normal.

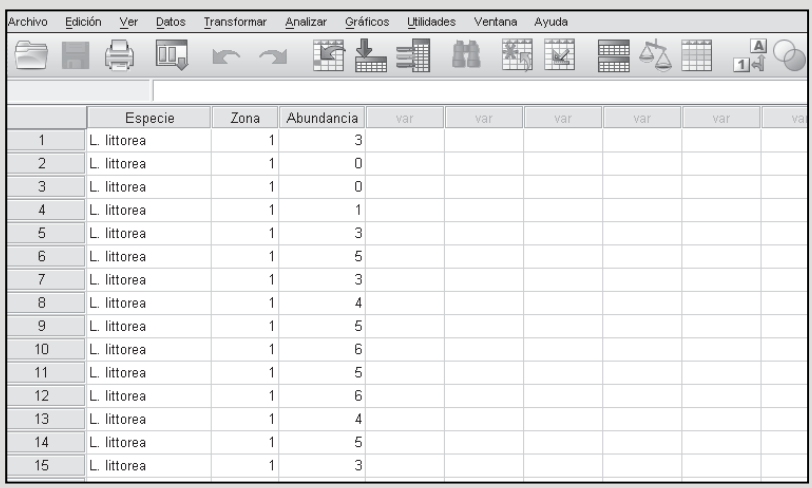
Además, es importante resaltar que con la estimación con R, y también veremos más adelante que con la estimación con SPSS, para *L. saxatilis* en la zona 1 $p = 0,001$ en el test de K-S y, por tanto, también se rechazaría la hipótesis de que la especie tiene una distribución Normal.

CUADRO VI.2. *(Continuación)*[SPSS]

Aplicación con SPSS.

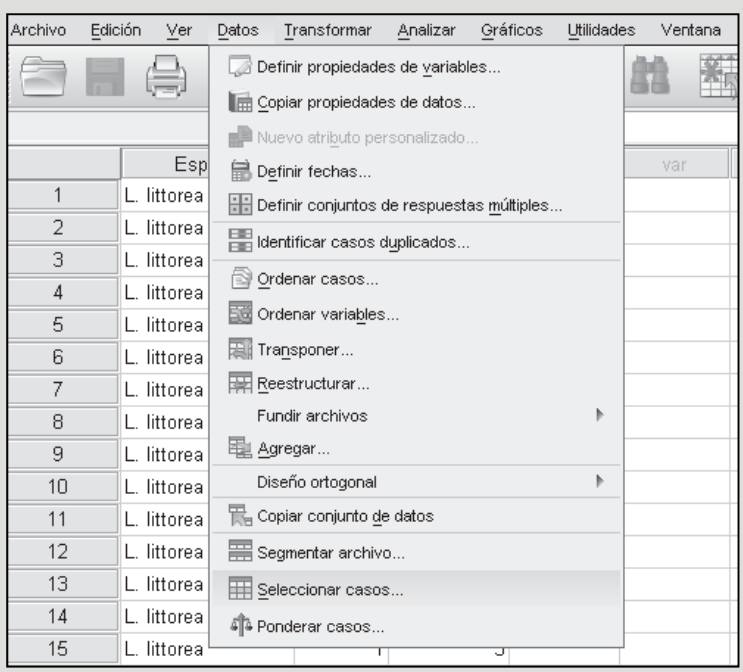
EJEMPLO. Vamos a utilizar el mismo caso explicado en la aplicación con R, el cual se encuentra en el archivo **Cuadro VI.1.sav**.

Paso 1. En SPSS los datos hay que organizarlos como se muestra en el archivo y en la siguiente ventana.



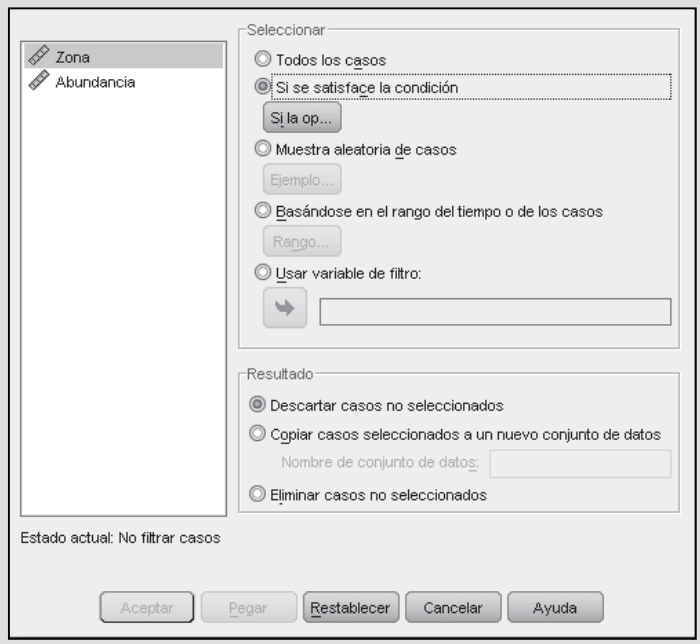
	Especie	Zona	Abundancia
1	L. littorea	1	3
2	L. littorea	1	0
3	L. littorea	1	0
4	L. littorea	1	1
5	L. littorea	1	3
6	L. littorea	1	5
7	L. littorea	1	3
8	L. littorea	1	4
9	L. littorea	1	5
10	L. littorea	1	6
11	L. littorea	1	5
12	L. littorea	1	6
13	L. littorea	1	4
14	L. littorea	1	5
15	L. littorea	1	3

Paso 2. En primer lugar hay que seleccionar el grupo que vamos a analizar. Para ello se elige en el menú «Datos» ⇨ «Seleccionar casos…».

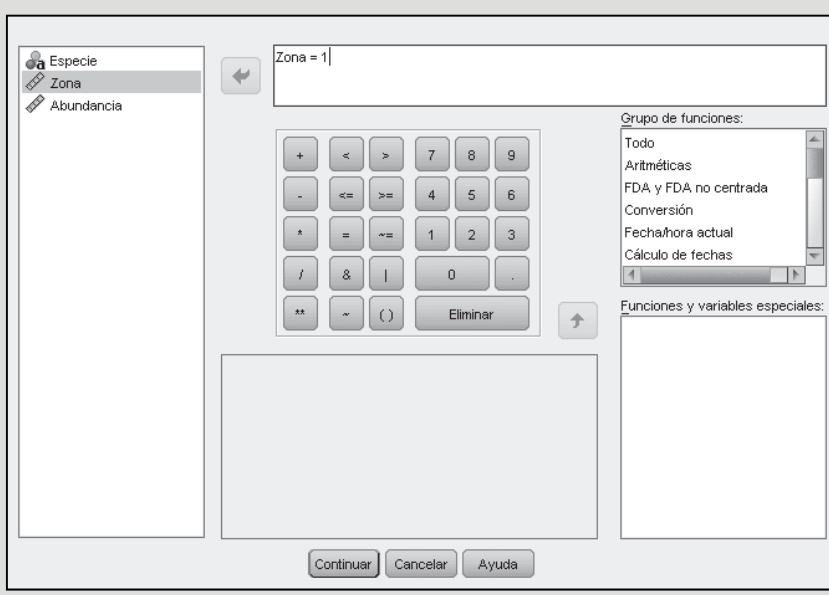


CUADRO VI.2. *(Continuación)*[SPSS]

Paso 3. En la ventana que aparece se marca «Si satisface la condición» y se pulsa el botón «Si la op…».

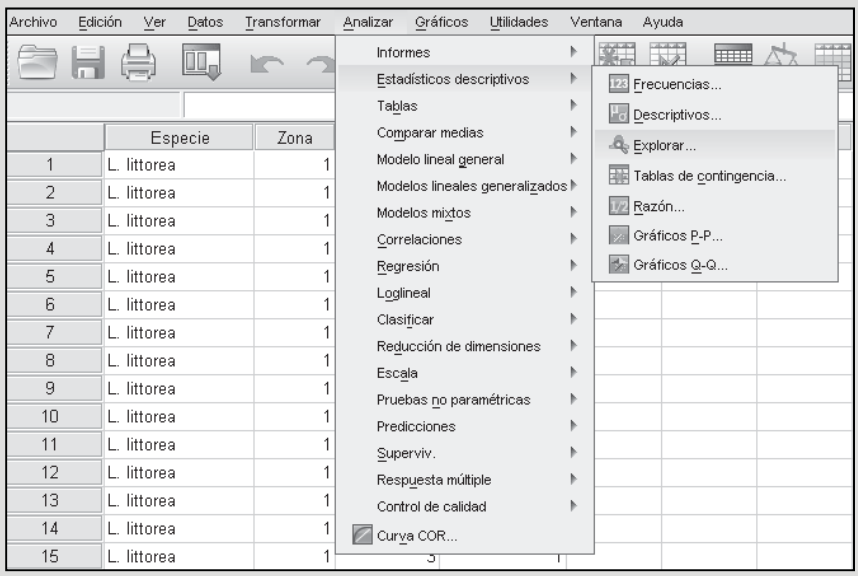


Paso 4. En la ventana que aparece se seleccionan los grupos con los que se va a trabajar. En este ejemplo seleccionamos «Zona=1», así podremos determinar si la abundancia de las tres especies presenta una distribución Normal en la zona 1.

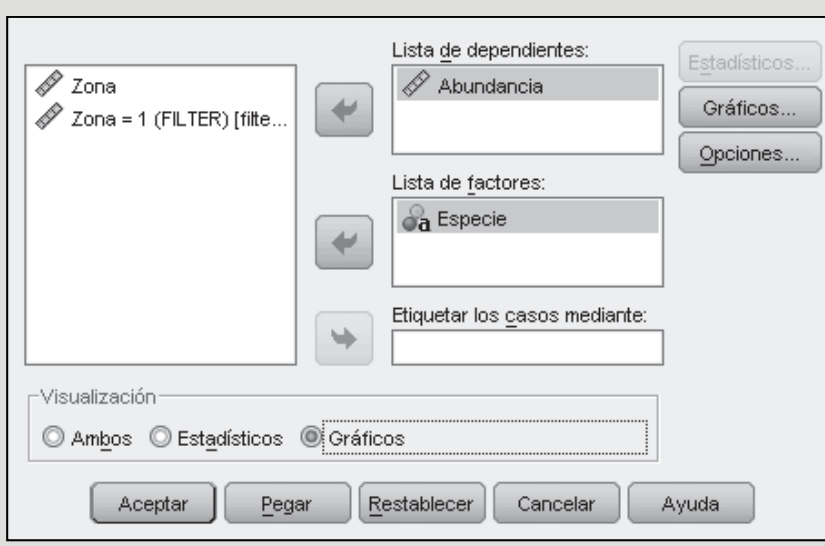


CUADRO VI.2. *(Continuación)*[SPSS]

Paso 5. A continuación se selecciona «Analizar» ⇨ «Estadísticos descriptivos…» ⇨ «Explorar…».

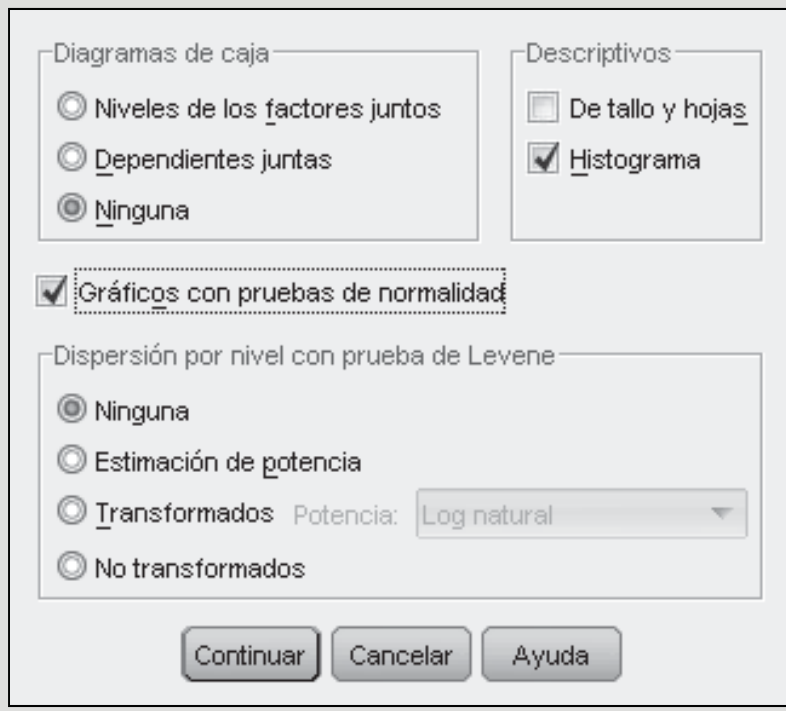


Paso 6. En la ventana que aparece, en «Lista de dependientes» se incluyen las variables de las que se va a determinar si presentan una distribución Normal, en nuestro ejemplo «Abundancia». En «Lista de factores» se incluye «Especie», con lo cual podemos determinar la distribución para las tres especies a la vez en la zona 1. En «Visualización» se selecciona «Gráficos» y, por último se pulsa el botón «Gráficos».



CUADRO VI.2. *(Continuación)*[SPSS]

Paso 7. A continuación en «Diagrama de cajas» se selecciona «Ninguna». En «Descriptivos» se selecciona «Histograma» y se marca «Gráficos con pruebas de normalidad».



Paso 8. Los resultados se muestran en la siguiente tabla. Para las 3 especies, las probabilidades del test K-S fueron menores de 0,05 (p = 0,001 para *Littorina littorea,* p = 0,001 para *Littorina saxatilis* y $p < 0,001$ para *Nucella lapillus*), por tanto se rechaza la hipótesis de que las distribuciones sean Normales. El test de Shapiro-Wilk corroboró los resultados del test K-S para *L. littorea* (p = 0,002) y *N. lapillus* (p = 0,006), pero en el caso de *L. saxatilis* p es igual a 0,065, lo cual indicaría que se acepta que la distribución de esta especie es Normal en la zona 1. Cuando existan discrepancias entre dos pruebas hay que adoptar una posición conservadora y rechazar la hipótesis nula, es decir, se rechaza la hipótesis de que la distribución sea Normal.

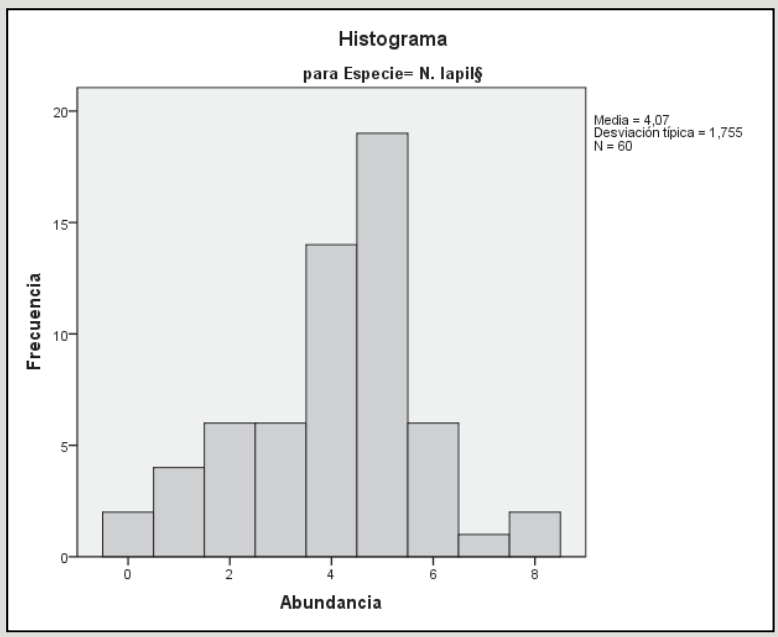
Pruebas de normalidad

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Especie	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Abundancia	L. litto	,184	41	,001	,901	41	,002
	L. saxat	,205	32	,001	,938	32	,065
	N. lapil	,185	60	,000	,941	60	,006

a. Corrección de la significación de Lilliefors

CUADRO VI.2. *(Continuación)*[SPSS]

Paso 9. A modo de ejemplo de los gráficos se muestran el histograma de frecuencias de *Nucella lapillus* y el gráfico Q-Q de *Littorina littorea*.



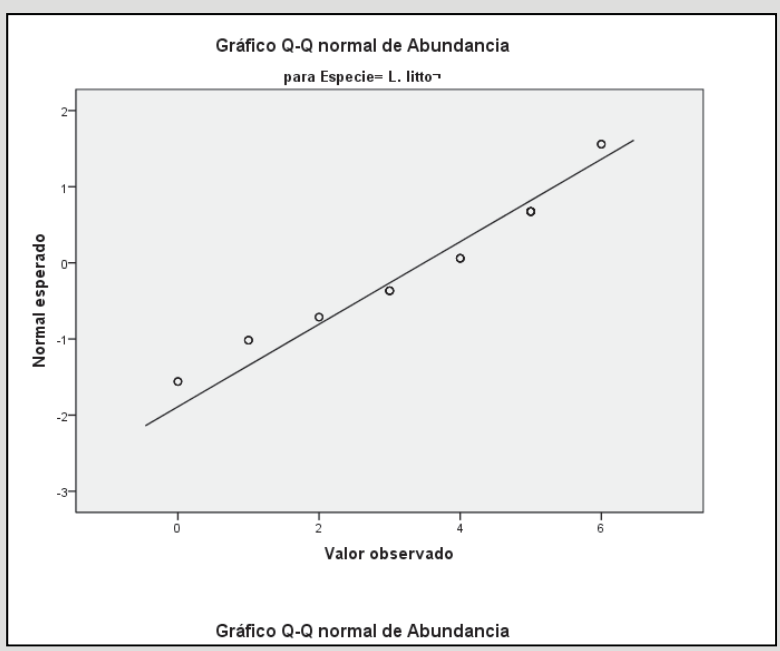


Gráfico Q-Q normal de Abundancia

Capítulo VII

CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUALITATIVAS

En este capítulo vamos a desarrollar los tratamientos estadísticos para realizar los contrastes de homogeneidad que sirven para trabajar con variables cualitativas que, como se describió en detalle al inicio del Capítulo III, son aquellas que no se expresan numéricamente y pueden ser dicotómicas (la variable solo puede tomar dos valores, como por ejemplo mujer o hombre) o politómicas (la variable puede tener tres o más valores), existiendo dentro de estas últimas dos tipos: nominales y ordinales.

En un contraste de homogeneidad lo que se pretende es comprobar si dos o más muestras pertenecen a una misma población o que dos muestras pertenecientes a una misma población no hayan sufrido alteraciones. En otras palabras, estamos interesados en conocer si los datos de mediciones de una variable en varios grupos o épocas tienen una misma distribución, o si cada medición es diferente. Podemos distinguir dos tipos de contrastes en función de la naturaleza de las variables, ya que se puede trabajar con mediciones independientes (por ejemplo, dos muestras elegidas en zonas distintas) o pareadas (por ejemplo, podemos medir la misma variable dos veces).

VII.1. Muestras independientes

VII.1.1. VARIABLES POLITÓMICAS

Las pruebas son el contraste Chi-cuadrado y la razón de verosimilitud (test *G*). Los datos se presentan en una tabla de contingencia (Tabla VII).

Tabla VII.1. Tabla de contingencia para la prueba de homogeneidad de muestras. *O* es la frecuencia observada, *k* es el número de categorías, *h* el número de muestras, y *n* el número de casos observados. Así, por ejemplo, O_{35} correspondería a la frecuencia observada de la muestra 3 y clase 5, n_{c3} al número total de frecuencias de la clase 3 y n_{m2} al número total de frecuencias de la muestra 2.

	Clase					
Muestra	***1***	***2***	***...***	***...***	***k***	**Total**
1	O_{11}	O_{12}	...	...	O_{1k}	$\mathbf{n_{m1}}$
2	O_{21}	O_{22}	...	...	O_{2k}	$\mathbf{n_{m2}}$
...	...	...	...	...	....	...
h	O_{h1}	O_{h2}	...	...	O_{hk}	$\boldsymbol{n_{mh}}$
Total	$\boldsymbol{n_{c1}}$	$\boldsymbol{n_{c2}}$	...	...	$\boldsymbol{n_{ck}}$	***n***

La tabla está formada por filas (h) y columnas (k) en las que se introduce la frecuencia de cada categoría analizada. Los grados de libertad de estas tablas son (h - 1)*(k - 1) que se corresponde con el número de casillas que pueden variar de forma independiente si se conoce las frecuencia total de cada fila y columna.

VII.1.1.1. Chi-cuadrado

La hipótesis nula de esta prueba supone que las muestras pertenecen a la misma población y, por tanto, la proporción de frecuencias esperadas es la misma en todas las muestras. Para contrastarla se calculan los valores de frecuencia esperados para cada celda y se comparan con las frecuencias observadas. En caso de que las diferencias sean pequeñas y, en consecuencia, el estadístico de contraste menor que un valor crítico, se acepta la hipótesis de que las muestras son homogéneas (tienen una misma distribución). La fórmula para el cálculo de los valores esperados es:

$$E_{ij} = \frac{n_{mi} n_{cj}}{n}$$

Una vez obtenidos los valores esperados se calcula el estadístico de contraste χ^2 con la siguiente fórmula:

$$\chi^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Se busca en la Tabla χ^2 (Tabla 3 del Apéndice) el valor crítico para (h-1)*(k-1) grados de libertad y para el nivel de significación elegido (generalmente 0,05). Si χ^2 crítico es mayor que χ^2 calculado se acepta la hipótesis nula de que son homogéneas, mientras que si χ^2 crítico es menor que χ^2 calculado, (o $p <$ nivel de significación), la hipótesis se rechaza. Para la aplicación de esta prueba se requiere que las frecuencias esperadas no tomen valores menores que 5. En el Cuadro VII.1 se indica la forma de realizar el test χ^2.

VII.1.1.2 Razón de verosimilitud (test *G*)

Este contraste es muy similar al test χ^2, ya que también cuantifica diferencias entre valores esperados y observados. Además, los grados de libertad se calculan del mismo modo que en el test χ^2 de homogeneidad de muestras y el valor crítico del test se obtiene a partir de la Tabla χ^2 (Tabla 3 del Apéndice).

La tabla de contingencia (Tabla VII.1) y el cálculo de los valores esperados son los mismos que los utilizados en el test χ^2 de homogeneidad. La fórmula del estadístico de contraste es:

$$G = 2 \sum_{m=1}^{h} \sum_{c=1}^{k} O_{mc} \ln \frac{O_{mc}}{E_{mc}}$$

Del mismo modo que en el contraste anterior, se busca en la Tabla χ^2 (Tabla 3 del Apéndice) el valor crítico con los grados de libertad y el nivel de significación

determinados. Si χ^2 crítico es mayor que *G* se acepta la hipótesis nula de que las muestras son homogéneas.

El estadístico G tiene una distribución aproximadamente χ^2. Existen algunas correcciones que mejoran la aproximación, como la de Williams (1976), aplicable cuando las diferentes categorías tienen frecuencias similares.

$$G_{adj} = \frac{G}{1 + \frac{a+1}{6n}}$$

donde *n* es el número de datos y *a* las categorías en las que están clasificados. Los pasos para su obtención se encuentran en el Cuadro VII.1.

CUADRO VII.1. Contraste de homogeneidad con χ^2 y test *G*

Aplicación con R.

EJEMPLO. En el archivo **Cuadro VII.1.V.csv** hay datos del grado en que fuman hombres y mujeres en distintos centros de trabajo. Las categorías utilizadas fueron: 1 (no fumador), 2 (entre 1 y 10 cigarrillos diarios), 3 (entre 11 y 20 cigarrillos diarios), 4 (entre 1 y 2 cajetillas diarias) y 5 (más de 2 cajetillas diarias). También hay información de si alguno de los padres de estos trabajadores son fumadores y las categorías son: trabajadores en que alguno de sus padres es fumador (categoría con valor 1) y el otro grupo son aquellos en que ninguno de sus padres es fumador (categoría con valor 0). Se quiere determinar, en cada uno de los centros de trabajo, si hay diferencias entre el grupo de trabajadores con alguno de los padres fumadores y el grupo en que sus padres no fuman, en la proporción de los distintos tipos de fumadores.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro VII.1.R**, que se muestra en la siguiente ventana. Solo es necesario especificar lo que se solicita en la sección modificable por el usuario.

```
# Cuadro VII.1
# Contraste de homogeneidad con X^2 y test G.

#########################################################
# Sección modificable por el usuario
#########################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro VII.1.V.csv",encoding="latin1")

# Selección de las variables de interés
varInteres<-c("Grado.de.fumador","Padres.Fumadores")
#varInteres<-c("Sexo","Grado.de.fumador")

# Selección de las variables de agrupación.
#varAgrupa<-NULL
#varAgrupa<-c("Sexo")
#varAgrupa<-c("Sexo","Centro.de.trabajo")
varAgrupa<-c("Centro.de.trabajo")
```

CUADRO VII.1. *(Continuación)*[R]

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro VII.1.V.csv**.

```
datos<-read.csv2("Cuadro VII.1.V.csv",encoding="latin1")
```

Paso 2. Se seleccionan las dos variables que se quieren estudiar, en este caso si los padres son fumadores y el grado de fumador.

```
varInteres<-c("Grado.de.fumador","Padres.Fumadores")
#varInteres<-c("Sexo", "Grado.de.fumador")
```

Paso 3. Es posible hacer los cálculos estadísticos agrupando las variables de la forma que se quiera. En el caso de seleccionar «NULL», no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna. En el ejemplo se selecciona que los estadísticos se hagan a nivel de centro de trabajo, aunque se pueden seleccionar varias variables. Sin embargo, estas variables de agrupación nunca pueden coincidir con las variables de interés seleccionadas en el paso anterior.

```
#varAgrupa<-NULL
#varAgrupa<-c("Sexo")
#varAgrupa<-c("Sexo","Centro.de.trabajo")
varAgrupa<-c("Centro.de.trabajo")
```

Paso 4. A continuación es necesario especificar los tipos de contrastes que se quieren realizar. En nuestro ejemplo, la prueba χ^2, la prueba G sin y con corrección de Williams.

```
# 1. Prueba chi-cuadrado.
# 2. Prueba G sin corrección de Williams
# 3. Prueba G con corrección de Williams
# 4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2)
# 5. Prueba "exacta" de Fisher (Sólo para tablas 2x2)

pruebas<-c(1,2,3)
```

Paso 5. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
nomSalida<-"Salida Cuadro VII.1.V.txt"
```

Paso 6. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*.

Veamos como ejemplo el caso del centro de trabajo 1, que se muestra en la siguiente página. En primer lugar se muestra la tabla de contingencia. Posteriormente se observan los resultados de los contrastes. La probabilidad es menor que 0,05 en los 3 contrastes, p = 0,025 en el test χ^2, p = 0,021 en el test G sin la corrección y p = 0,024 con la corrección de Williams.

CUADRO VII.1. *(Continuación)*[R]

Por lo tanto, se rechaza la hipótesis nula de que las muestras son homogéneas, es decir, existen diferencias significativas entre el grupo con alguno de sus padres fumadores y el grupo sin padres fumadores, en el porcentaje de los distintos tipos de fumadores.

```
$`Centro trabajo 1`
$`Centro trabajo 1`$tabla
                              Padres.Fumadores
Grado.de.fumador               No Si
  1 a 10 cigarrillos diarios    7 13
  11 a 20 cigarrillos diarios 14 24
  De 1 a 2 cajetillas diarias 15 15
  Más de 2 cajetillas diarias 16 17
  No fuma                     19  6

$`Centro trabajo 1`$listaPruebas
$`Centro trabajo 1`$listaPruebas[[1]]

        Pearson's Chi-squared test

data:  tbl1
X-squared = 11.1206, df = 4, p-value = 0.02524


$`Centro trabajo 1`$listaPruebas[[2]]

        Log likelihood ratio (G-test) test of independence without correction

data:  tbl1
Log likelihood ratio statistic (G) = 11.5149, X-squared df = 4, p-value
= 0.02135


$`Centro trabajo 1`$listaPruebas[[3]]

        Log likelihood ratio (G-test) test of independence with Williams'
        correction

data:  tbl1
Log likelihood ratio statistic (G) = 11.2708, X-squared df = 4, p-value
= 0.02368
```

En el centro de trabajo 2 la probabilidad es también menor que 0,05 en los 3 contrastes, $p < 0{,}001$ en el test χ^2, $p < 0{,}001$ en el test G sin la corrección y $p < 0{,}001$ con la corrección de Williams. Por lo tanto, también se rechaza la hipótesis nula de que las muestras son homogéneas y la conclusión es la misma que la obtenida en el centro de trabajo 1.

```
$`Centro trabajo 2`
$`Centro trabajo 2`$tabla
                              Padres.Fumadores
Grado.de.fumador               No Si
  1 a 10 cigarrillos diarios    7 18
  11 a 20 cigarrillos diarios   3 24
  De 1 a 2 cajetillas diarias   6 17
  Más de 2 cajetillas diarias   2 12
  No fuma                      15  7

$`Centro trabajo 2`$listaPruebas
$`Centro trabajo 2`$listaPruebas[[1]]

        Pearson's Chi-squared test

data:  tbl1
X-squared = 21.8309, df = 4, p-value = 0.0002166


$`Centro trabajo 2`$listaPruebas[[2]]

        Log likelihood ratio (G-test) test of independence without correction

data:  tbl1
Log likelihood ratio statistic (G) = 21.2078, X-squared df = 4, p-value
= 0.000288
```

CUADRO VII.1. *(Continuación)*[R]

En los otros centros de trabajo también se rechaza la hipótesis nula:

Centro de trabajo 3

χ^2 = 19,8, *df* = 4, *p* < 0,001

Test de la G = 20,2, *df* = 4, *p* < 0,001

Test de la G con la corrección de Williams = 19,6, *df* = 4, *p* < 0,001

Centro de trabajo 4

χ^2 = 20,04, *df* = 4, *p* < 0,001

Test de la G = 19,5, *df* = 4, *p* < 0,001

Test de la G con la corrección de Williams = 19,07, *df* = 4, *p* < 0,001

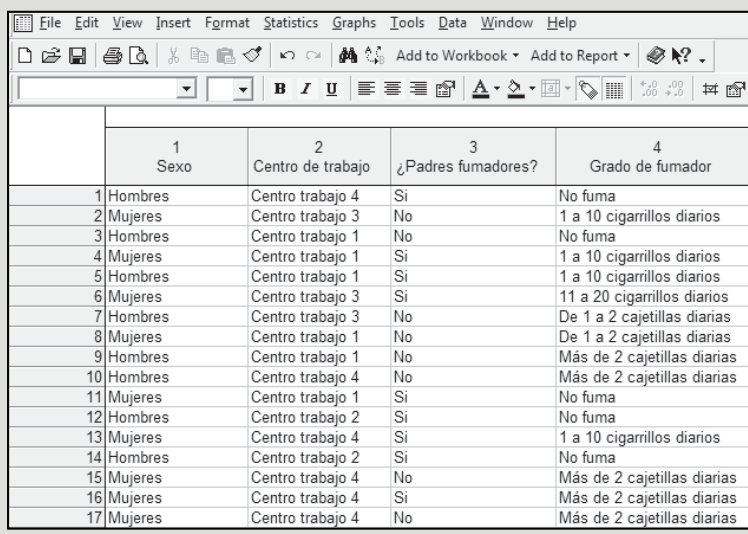
```
$`Centro trabajo 3`
$`Centro trabajo 3`$tabla
                          Padres.Fumadores
Grado.de.fumador           No Si
  1 a 10 cigarrillos diarios  11  8
  11 a 20 cigarrillos diarios 10 27
  De 1 a 2 cajetillas diarias 10 18
  Más de 2 cajetillas diarias  4 23
  No fuma                     10  3

$`Centro trabajo 3`$listaPruebas
$`Centro trabajo 3`$listaPruebas[[1]]

        Pearson's Chi-squared test

data:  tbl1
X-squared = 19.882, df = 4, p-value = 0.0005269


$`Centro trabajo 3`$listaPruebas[[2]]

        Log likelihood ratio (G-test) test of independence without correction

data:  tbl1
Log likelihood ratio statistic (G) = 20.2167, X-squared df = 4, p-value
= 0.0004525


$`Centro trabajo 3`$listaPruebas[[3]]

        Log likelihood ratio (G-test) test of independence with Williams'
        correction

data:  tbl1
Log likelihood ratio statistic (G) = 19.6154, X-squared df = 4, p-value
= 0.0005947


$`Centro trabajo 4`
$`Centro trabajo 4`$tabla
                          Padres.Fumadores
Grado.de.fumador           No Si
  1 a 10 cigarrillos diarios  11 24
  11 a 20 cigarrillos diarios 14 30
  De 1 a 2 cajetillas diarias  4 18
  Más de 2 cajetillas diarias  8 26
  No fuma                     19  8

$`Centro trabajo 4`$listaPruebas
$`Centro trabajo 4`$listaPruebas[[1]]

        Pearson's Chi-squared test

data:  tbl1
X-squared = 20.044, df = 4, p-value = 0.0004895
```

CUADRO VII.1. *(Continuación)*[STATISTICA]

Aplicación con STATISTICA.

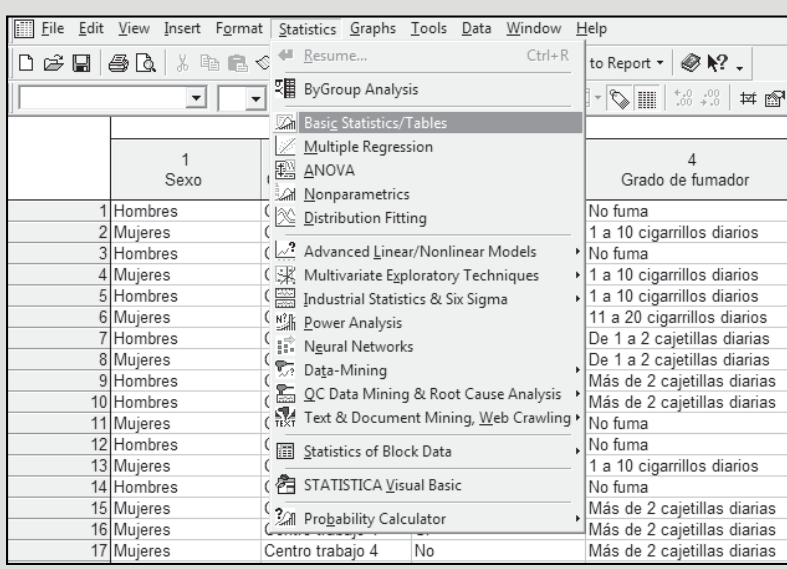
EJEMPLO. Vamos a utilizar los mismos datos que en el caso de la estimación con R, los cuales se encuentran en el archivo **Cuadro VII.1.sta**. Sin embargo, en este caso se quiere determinar, en cada uno de los centros de trabajo, si hay diferencias entre hombres y mujeres en la proporción de los distintos tipos de fumadores.

Paso 1. En STATISTICA los datos hay que organizarlos en columnas como se muestra en la siguiente ventana.



	1 Sexo	2 Centro de trabajo	3 ¿Padres fumadores?	4 Grado de fumador
1	Hombres	Centro trabajo 4	Si	No fuma
2	Mujeres	Centro trabajo 3	No	1 a 10 cigarrillos diarios
3	Hombres	Centro trabajo 1	No	No fuma
4	Mujeres	Centro trabajo 1	Si	1 a 10 cigarrillos diarios
5	Hombres	Centro trabajo 1	Si	1 a 10 cigarrillos diarios
6	Mujeres	Centro trabajo 3	Si	11 a 20 cigarrillos diarios
7	Hombres	Centro trabajo 3	No	De 1 a 2 cajetillas diarias
8	Mujeres	Centro trabajo 1	No	De 1 a 2 cajetillas diarias
9	Hombres	Centro trabajo 1	No	Más de 2 cajetillas diarias
10	Hombres	Centro trabajo 4	No	Más de 2 cajetillas diarias
11	Mujeres	Centro trabajo 1	Si	No fuma
12	Hombres	Centro trabajo 2	Si	No fuma
13	Mujeres	Centro trabajo 4	Si	1 a 10 cigarrillos diarios
14	Hombres	Centro trabajo 2	Si	No fuma
15	Mujeres	Centro trabajo 4	No	Más de 2 cajetillas diarias
16	Mujeres	Centro trabajo 4	Si	Más de 2 cajetillas diarias
17	Mujeres	Centro trabajo 4	No	Más de 2 cajetillas diarias

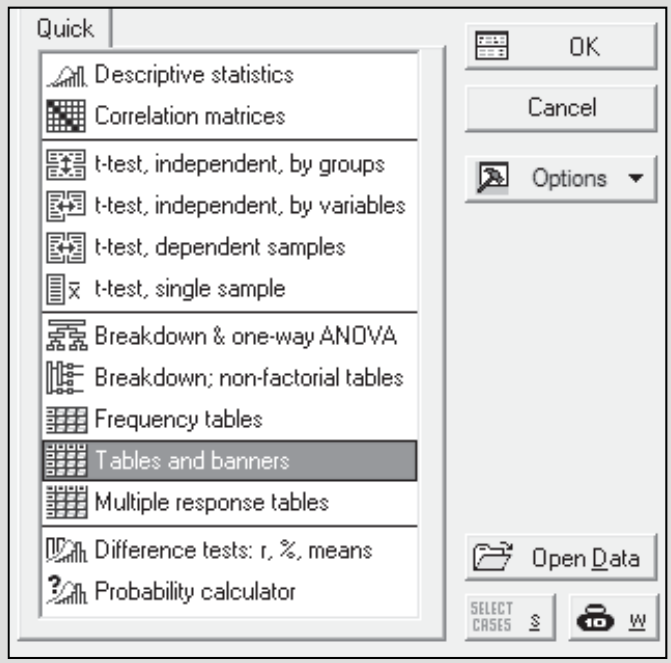
Paso 2. Se selecciona «Statistics» ⇨ «Basic Statistics/Tables».



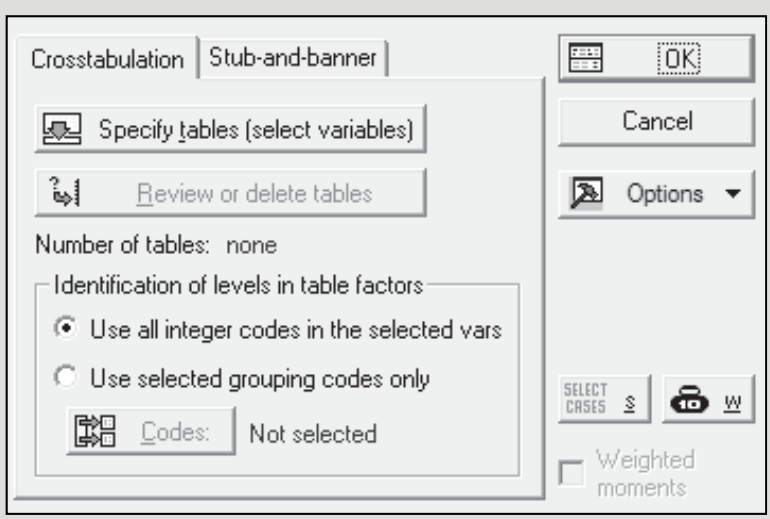
	1 Sexo		4 Grado de fumador
1	Hombres		No fuma
2	Mujeres		1 a 10 cigarrillos diarios
3	Hombres		No fuma
4	Mujeres		1 a 10 cigarrillos diarios
5	Hombres		1 a 10 cigarrillos diarios
6	Mujeres		11 a 20 cigarrillos diarios
7	Hombres		De 1 a 2 cajetillas diarias
8	Mujeres		De 1 a 2 cajetillas diarias
9	Hombres		Más de 2 cajetillas diarias
10	Hombres		Más de 2 cajetillas diarias
11	Mujeres		No fuma
12	Hombres		No fuma
13	Mujeres		1 a 10 cigarrillos diarios
14	Hombres		No fuma
15	Mujeres		Más de 2 cajetillas diarias
16	Mujeres		Más de 2 cajetillas diarias
17	Mujeres	Centro trabajo 4	No	Más de 2 cajetillas diarias

CUADRO VII.1. *(Continuación)*[STATISTICA]

Paso 3. En la ventana que aparece se selecciona «Tables and banners».

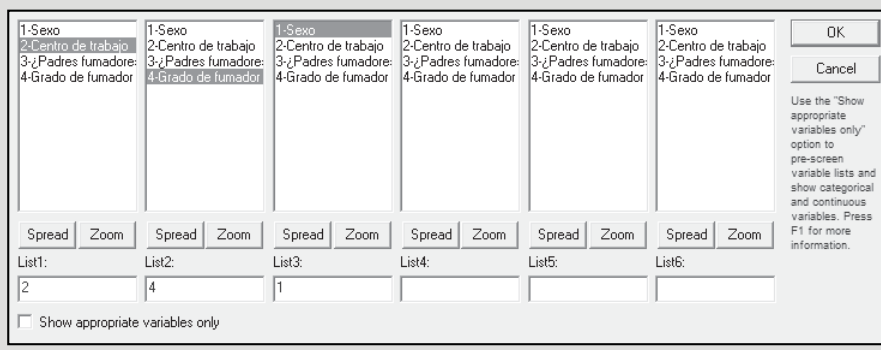


Paso 4. Activamos la pestaña «Crosstabulation» (tabla de contingencia) donde se nos permite seleccionar las categorías con las que queremos trabajar («Use selected grouping codes only») o usar todas las categorías de los datos («Use all integer codes in the selected vars»), que es la opción que trae por defecto el programa. Nosotros trabajaremos con todas las categorías y no modificaremos este punto.

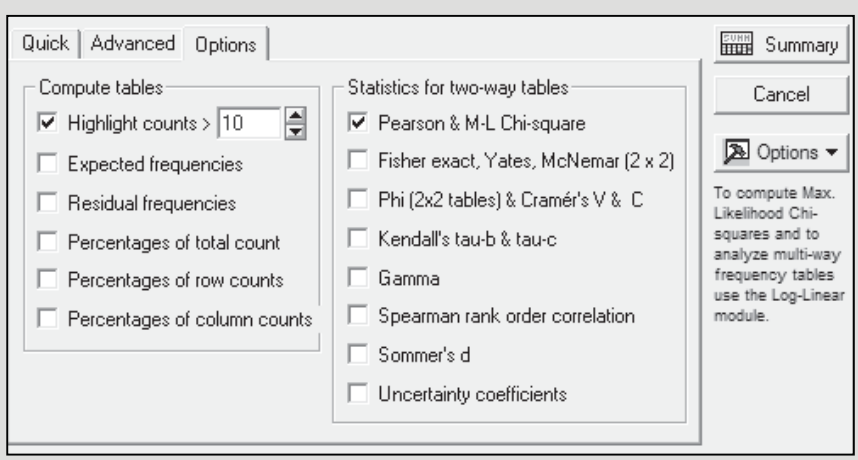


CUADRO VII.1. *(Continuación)* STATISTICA

Paso 5. En la ventana anterior pulsamos sobre «Specify tables (select variables)» y aparece la ventana que se muestra a continuación, donde se pueden seleccionar hasta 6 variables distintas. Es importante mencionar que la última variable que se especifique será la que define los grupos a comparar, en este caso se quiere comparar entre hombres y mujeres y, por ello, se selecciona «Sexo» como última variable. La primera variable se usa para diferenciar si se quiere hacer el análisis por grupos. Como en este caso queremos ver las diferencias para cada uno de los centros de trabajo, introducimos esta variable primero. Por último, como lo que se pretende es comparar entre hombres y mujeres la proporción de los distintos tipos de fumadores, esta variable («Grado de fumador») se pone justo antes de la variable «Sexo».

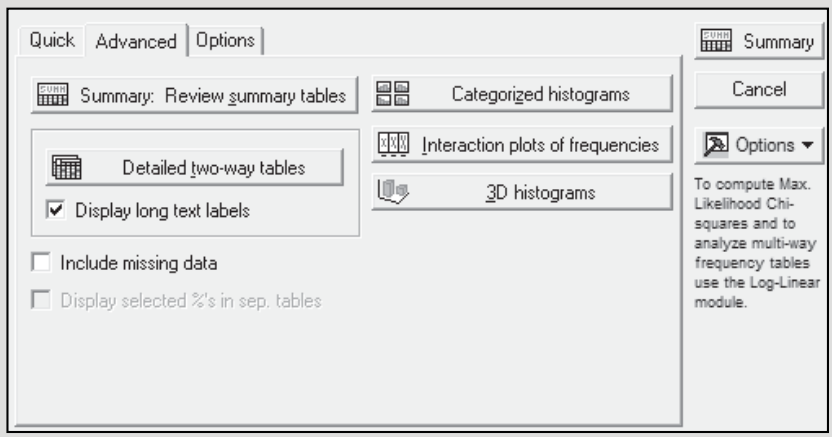


Paso 6. En la ventana del paso 4 pulsamos en «OK» y nos aparece la ventana que está a continuación, donde hay tres pestañas diferentes, de las cuales dos son de resultados, «Quick» y «Advanced», y una para seleccionar los estadísticos «Options». En la pestaña «Options» seleccionamos la opción «Pearson & ML Chi-square». Estos 2 estadísticos se corresponden con el test χ^2 (Pearson Chi-square) y el test *G* (Maximum Likelihood Chi-square), respectivamente.



CUADRO VII.1. *(Continuación)*[STATISTICA]

Paso 7. En la pestaña "Advanced" aparece la siguiente ventana, donde pulsando el icono "Summary: Review summary tables" aparece la tabla de contingencia de nuestros resultados.



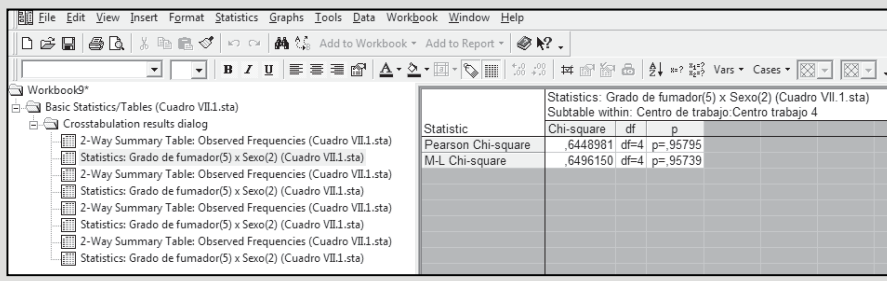
Summary Frequency Table (Cuadro VII.1.sta)
Marked cells have counts > 10
(Marginal summaries are not marked)

Centro de trabajo	Grado de fumador	Sexo Hombres	Sexo Mujeres	Row Totals
Centro trabajo 4	No fuma	9	18	27
Centro trabajo 4	1 a 10 cigarrillos diarios	14	21	35
Centro trabajo 4	11 a 20 cigarrillos diarios	18	26	44
Centro trabajo 4	De 1 a 2 cajetillas diarias	9	13	22
Centro trabajo 4	Más de 2 cajetillas diarias	12	22	34
Total		62	100	162
Centro trabajo 3	No fuma	6	7	13
Centro trabajo 3	1 a 10 cigarrillos diarios	7	12	19
Centro trabajo 3	11 a 20 cigarrillos diarios	13	24	37
Centro trabajo 3	De 1 a 2 cajetillas diarias	5	23	28
Centro trabajo 3	Más de 2 cajetillas diarias	6	21	27
Total		37	87	124
Centro trabajo 1	No fuma	7	18	25
Centro trabajo 1	1 a 10 cigarrillos diarios	7	13	20
Centro trabajo 1	11 a 20 cigarrillos diarios	13	25	38
Centro trabajo 1	De 1 a 2 cajetillas diarias	8	22	30
Centro trabajo 1	Más de 2 cajetillas diarias	14	19	33
Total		49	97	146
Centro trabajo 2	No fuma	10	12	22
Centro trabajo 2	1 a 10 cigarrillos diarios	7	18	25
Centro trabajo 2	11 a 20 cigarrillos diarios	13	14	27
Centro trabajo 2	De 1 a 2 cajetillas diarias	8	15	23
Centro trabajo 2	Más de 2 cajetillas diarias	8	6	14
Total		46	65	111
Column Total		194	349	543

CUADRO VII.1. *(Continuación)*[STATISTICA]

Paso 8. En la ventana de la pestaña «Advanced» seleccionamos el icono «Detailed two-ways table» para que el programa nos muestre el resultado de los estadísticos solicitados.

Aparece una ventana donde a la izquierda podemos seleccionar los resultados obtenidos para cada centro de trabajo.



En la tabla de la derecha aparece una primera fila que contiene el estadístico χ^2 y una segunda con el test *G.* En ambos casos, se nos indica el valor del estadístico, los grados de libertad y la probabilidad del contraste. Por ejemplo, en el caso del centro de trabajo 1, la probabilidad es mayor de 0,05 para los 2 contrastes, p = 0,704 en el test χ^2 y p = 0,705 en el test *G*. Por lo tanto, se acepta la hipótesis nula de que las muestras son homogéneas, es decir, no existen diferencias significativas entre hombres y mujeres en el porcentaje de los distintos tipos de fumadores.

	Statistics: Grado de fumador(5) x Sexo(2) (Cuadro VII.1.sta) Subtable within: Centro de trabajo:Centro trabajo 1		
Statistic	Chi-square	df	p
Pearson Chi-square	2,174610	df=4	p=,70368
M-L Chi-square	2,169086	df=4	p=,70469

En el centro de trabajo 2 tampoco hay diferencias significativas entre hombres y mujeres en el porcentaje de los distintos tipos de fumadores con el test χ^2 (χ^2_4 = 4,35, p = 0,361) y el test *G* (test G = 4,41, *df* = 4, p = 0,353).

	Statistics: Grado de fumador(5) x Sexo(2) (Cuadro VII.1.sta) Subtable within: Centro de trabajo:Centro trabajo 2		
Statistic	Chi-square	df	p
Pearson Chi-square	4,350211	df=4	p=,36069
M-L Chi-square	4,411787	df=4	p=,35314

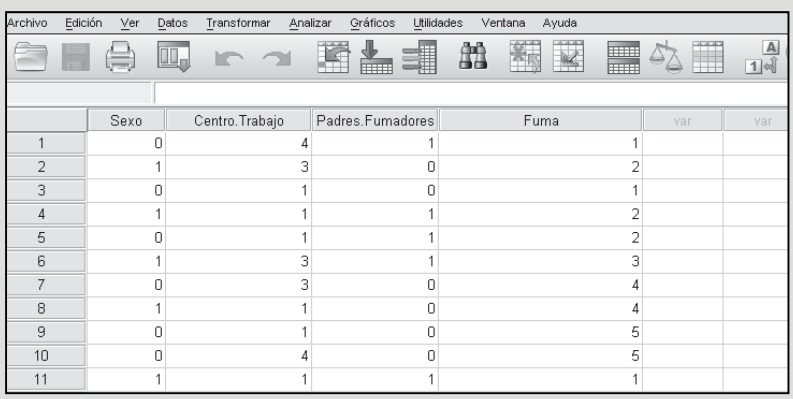
Los mismos resultados se obtienen para el centro de trabajo 3 (χ^2_4 = 5,26, p = 0,261 y test G = 5,35, *df* = 4, p = 0,253) y en el centro de trabajo 4 (χ^2_4 = 0,65, p = 0,958 y el test G = 0,65, *df* = 4, p = 0,957).

CUADRO VII.1. *(Continuación)*[SPSS]

Aplicación con SPPS.

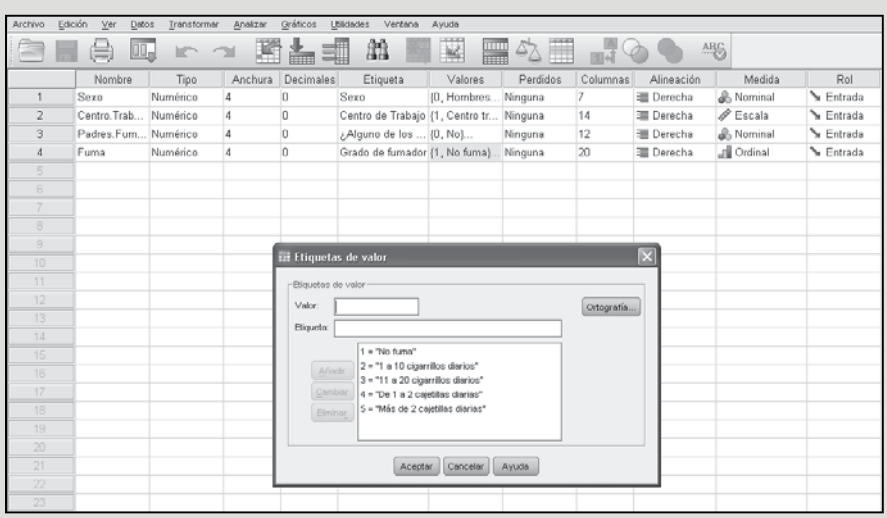
EJEMPLO. Vamos a utilizar los mismos datos que en el caso de la estimación con R y con STATISTICA, los cuales se encuentran en el archivo **Cuadro VII.1.sav**. Sin embargo, en este caso se quiere determinar si hay diferencias entre los centros de trabajo en la proporción de los distintos tipos de fumadores, haciendo la diferencia entre los trabajadores que tenían algún padre fumador y los que no tenían ningún padre fumador.

Paso 1. En SPSS los datos hay que organizarlos como se muestra en el archivo y en la siguiente ventana.


Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana Ayuda

	Sexo	Centro.Trabajo	Padres.Fumadores	Fuma	var	var
1	0	4	1	1		
2	1	3	0	2		
3	0	1	0	1		
4	1	1	1	2		
5	0	1	1	2		
6	1	3	1	3		
7	0	3	0	4		
8	1	1	0	4		
9	0	1	0	5		
10	0	4	0	5		
11	1	1	1	1		

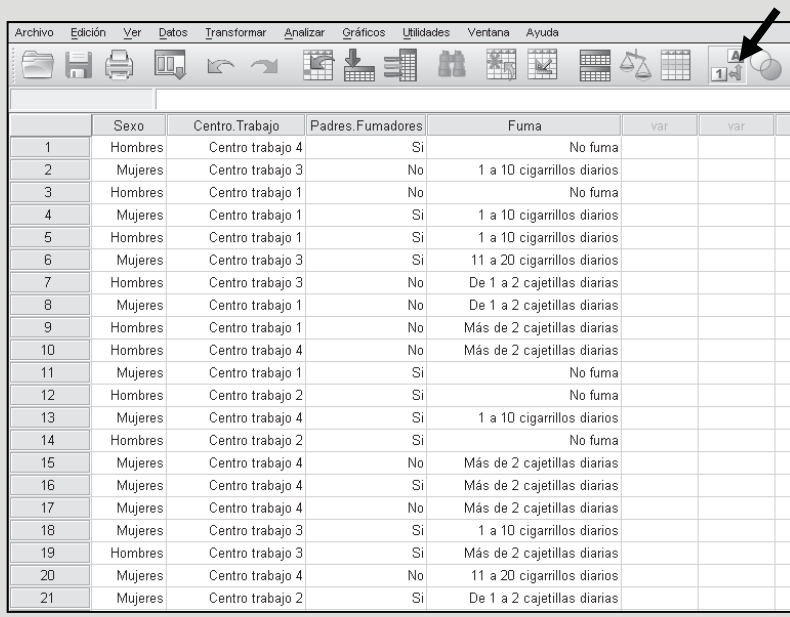
Paso 2. Al seleccionar la pestaña «Vista de variables» aparece la siguiente ventana donde es posible especificar el nombre de cada variable en «Nombre», el tipo de variable en «Tipo», etc. En «Valores» aparece la ventana que se muestra a continuación, donde es posible etiquetar a qué corresponde cada valor en cada una de las variables. Se muestra el ejemplo para la variable «Grado de fumador».



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Sexo	Numérico	4	0	Sexo	{0, Hombres...	Ninguna	7	Derecha	Nominal	Entrada
2	Centro.Trab...	Numérico	4	0	Centro de Trabajo	{1, Centro tr...	Ninguna	14	Derecha	Escala	Entrada
3	Padres.Fum...	Numérico	4	0	¿Alguno de los ...	{0, No}...	Ninguna	12	Derecha	Nominal	Entrada
4	Fuma	Numérico	4	0	Grado de fumador	{1, No fuma}...	Ninguna	20	Derecha	Ordinal	Entrada

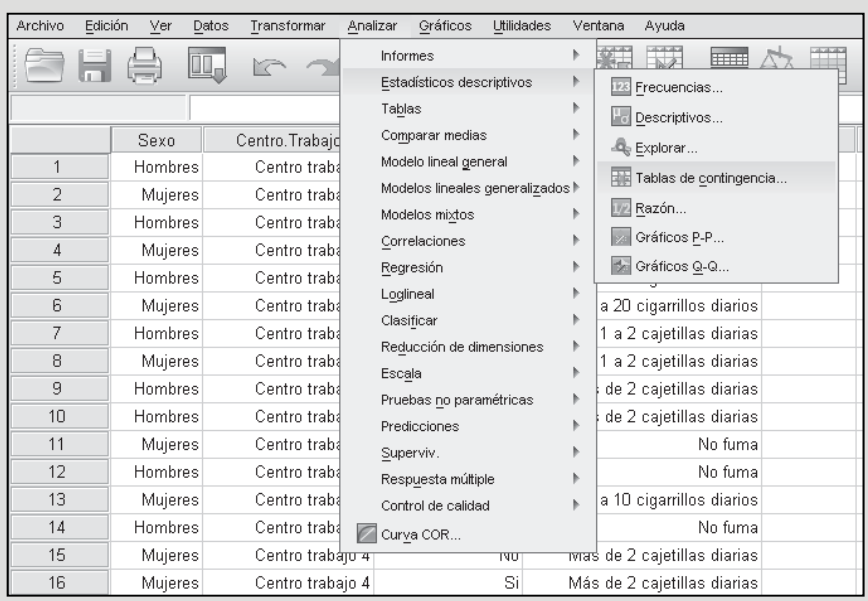
CUADRO VII.1. *(Continuación)*[SPSS]

Paso 3. Al volver a la ventana de la pestaña «Vista de datos», si pulsamos donde indica la flecha, aparecen –en vez de los números– las etiquetas que corresponden a cada valor.



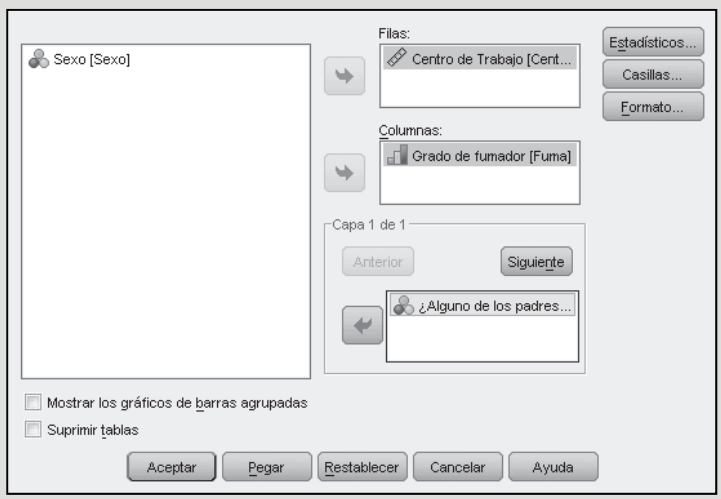
	Sexo	Centro.Trabajo	Padres.Fumadores	Fuma	var	var
1	Hombres	Centro trabajo 4	Si	No fuma		
2	Mujeres	Centro trabajo 3	No	1 a 10 cigarrillos diarios		
3	Hombres	Centro trabajo 1	No	No fuma		
4	Mujeres	Centro trabajo 1	Si	1 a 10 cigarrillos diarios		
5	Hombres	Centro trabajo 1	Si	1 a 10 cigarrillos diarios		
6	Mujeres	Centro trabajo 3	Si	11 a 20 cigarrillos diarios		
7	Hombres	Centro trabajo 3	No	De 1 a 2 cajetillas diarias		
8	Mujeres	Centro trabajo 1	No	De 1 a 2 cajetillas diarias		
9	Hombres	Centro trabajo 1	No	Más de 2 cajetillas diarias		
10	Hombres	Centro trabajo 4	No	Más de 2 cajetillas diarias		
11	Mujeres	Centro trabajo 1	Si	No fuma		
12	Hombres	Centro trabajo 2	Si	No fuma		
13	Mujeres	Centro trabajo 4	Si	1 a 10 cigarrillos diarios		
14	Hombres	Centro trabajo 2	Si	No fuma		
15	Mujeres	Centro trabajo 4	No	Más de 2 cajetillas diarias		
16	Mujeres	Centro trabajo 4	Si	Más de 2 cajetillas diarias		
17	Mujeres	Centro trabajo 4	No	Más de 2 cajetillas diarias		
18	Mujeres	Centro trabajo 3	Si	1 a 10 cigarrillos diarios		
19	Hombres	Centro trabajo 3	Si	Más de 2 cajetillas diarias		
20	Mujeres	Centro trabajo 4	No	11 a 20 cigarrillos diarios		
21	Mujeres	Centro trabajo 2	Si	De 1 a 2 cajetillas diarias		

Paso 4. A continuación se selecciona «Analizar» ⇨ «Estadísticos descriptivos» ⇨ «Tablas de contingencia…».

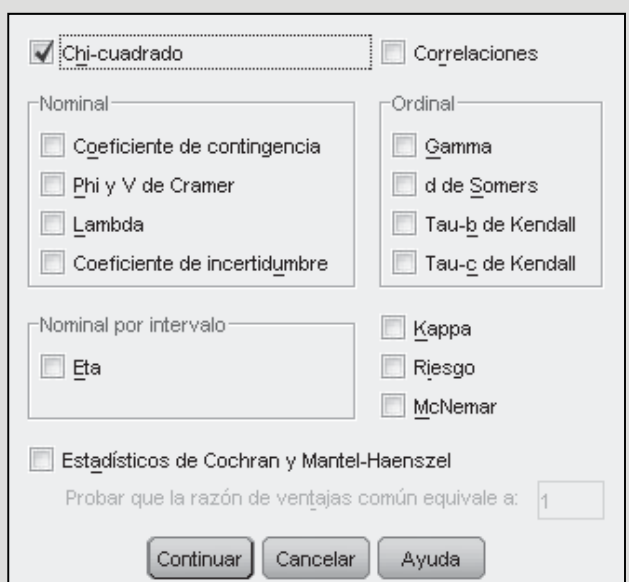


CUADRO VII.1. *(Continuación)*[SPSS]

Paso 5. En «Filas» hay que poner la variable que tiene los grupos a comparar, en este ejemplo «Centro de trabajo». En «Columnas» la variable que se quiere comparar, en este caso «Grado de fumador». Por último, como queremos hacer la distinción entre trabajadores con padres fumadores o no fumadores, en «Capa» se pone la variable «¿Alguno de los padres fuman?». Es posible hacer varias capas, es decir, diferentes tipos de agrupaciones.



Paso 6. En la ventana anterior hay que pulsar en «Estadísticos…» y en la ventana que aparece marcar «Chi-cuadrado».



CUADRO VII.1. *(Continuación)*[SPSS]

Paso 7. En la tabla de contingencia aparecen las frecuencias de todas las clases.

Tabla de contingencia Centro de Trabajo * Grado de fumador * ¿Alguno de los padres fuman?

¿Alguno de los padres fuman?			Grado de fumador					Total
			No fuma	1 a 10 cigarrillos diarios	11 a 20 cigarrillos diarios	De 1 a 2 cajetillas diarias	Más de 2 cajetillas diarias	
No	Centro de Trabajo	Centro trabajo 1	19	7	14	15	16	71
		Centro trabajo 2	15	7	3	6	2	33
		Centro trabajo 3	10	11	10	10	4	45
		Centro trabajo 4	19	11	14	4	8	56
	Total		63	36	41	35	30	205
Si	Centro de Trabajo	Centro trabajo 1	6	13	24	15	17	75
		Centro trabajo 2	7	18	24	17	12	78
		Centro trabajo 3	3	8	27	18	23	79
		Centro trabajo 4	8	24	30	18	26	106
	Total		24	63	105	68	78	338

Paso 8. Los resultados se muestran en la siguiente tabla. En el grupo de trabajadores que sus padres no fumaban, existen diferencias significativas en la proporción de los distintos tipos de fumadores entre centros de trabajo, tanto en el test χ^2 ($\chi^2_{12} = 21{,}06$, $p = 0{,}049$) como en el test G ($G = 22{,}65$, $df = 12$, $p = 0{,}031$). Por el contrario, en el grupo de trabajadores que alguno de sus padres fuma, no existen diferencias significativas en la proporción de los distintos tipos de fumadores entre centros de trabajo, tanto en el test χ^2 ($\chi^2_{12} = 11{,}38$, $p = 0{,}497$) como en el test G ($G = 12{,}28$, $df = 12$, $p = 0{,}424$).

Pruebas de chi-cuadrado

¿Alguno de los padres fuman?		Valor	gl	Sig. asintótica (bilateral)
No	Chi-cuadrado de Pearson	21,063[a]	12	,049
	Razón de verosimilitudes	22,653	12	,031
	Asociación lineal por lineal	3,212	1	,073
	N de casos válidos	205		
Si	Chi-cuadrado de Pearson	11,377[b]	12	,497
	Razón de verosimilitudes	12,279	12	,424
	Asociación lineal por lineal	,290	1	,590
	N de casos válidos	338		

a. 1 casillas (5,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 4,83.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 5,33.

VII.1.2. VARIABLES DICOTÓMICAS

Cuando ambas variables son dicotómicas, es decir tienen dos únicas categorías como, por ejemplo, presencia/ausencia o mujer/hombre, entonces tenemos una tabla de contingencia de 2x2 (Tabla VII.2), y se usa una corrección del test χ^2, la corrección de Yates (1934) o el test de Fisher (1922).

Tabla VII.2. Tabla de contingencia 2x2.

Variable 1	**Variable 2**		
	1	***2***	
			Total
1	*a*	*c*	a+c
2	*b*	*d*	b+d
Total	a+b	c+d	*n*

VII.1.2.1. Corrección de Yates

La corrección de Yates se aplica cuando una tabla de contingencia es 2x2 (véase Tabla VII.2), es decir, dos filas y dos columnas, y la muestra tiene un número pequeño de casos (con alguna frecuencia inferior a 5) ya que no es posible agrupar categorías.

En esta situación, el valor de χ^2 se sobreestima y, para corregirlo, se aplica la denominada corrección de Yates que modifica el cálculo del estadístico de la siguiente manera:

$$\chi^2 = \sum_{m=1}^{j} \sum_{c=1}^{i} \frac{(|O_{mc} - E_{mc}| - \frac{1}{2})^2}{E_{mc}}$$

En este caso el valor crítico de contraste sería el χ^2 para un grado de libertad. Si el valor obtenido es menor al valor tabulado de contraste, se acepta la hipótesis nula de que las muestras son homogéneas.

No obstante, conviene mencionar que la utilización de la corrección de Yates no exime de ciertos requerimientos acerca del tamaño muestral necesario para la utilización del estadístico χ^2. Como norma general, se exigirá que el 80% de las celdas en una tabla de contingencia tengan valores esperados mayores que 5. Así, en una tabla 2x2 será necesario que todas las celdas verifiquen esta condición, si bien en la práctica suele permitirse que una de ellas muestre frecuencias esperadas ligeramente por debajo de este valor.

En el Cuadro VII.2 se explica la forma de aplicar está corrección del test χ^2.

VII.1.2.2. Prueba de Fisher

La prueba de Fisher se utiliza con tablas de contingencia 2x2 y es la más aconsejable cuando exista una o más celdas con frecuencias esperadas menores que 5.

La prueba se basa en la determinación, a partir de una distribución hipergeométrica, de la probabilidad de obtener las frecuencias observadas cuando los totales de filas y columnas son constantes, teniendo como hipótesis nula que las variables son independientes. Su cálculo se explica en el Cuadro VII.2.

CUADRO VII.2. Corrección de Yates y prueba de Fisher

Aplicación con R.

EJEMPLO. En el archivo **Cuadro VII.2.V.csv** hay datos de la presencia del virus del herpes labial (0 es no portador y 1 es portador) en hombres y mujeres (1 es hombre y 2 es mujer). Se quiere determinar si el número de portadores del virus es diferente entre hombres y mujeres.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro VII.2.R**, que se muestra en la siguiente ventana. Solo es necesario introducir los datos que se solicitan en la sección modificable por el usuario.

```
# Cuadro VII.2
# Contraste de homogeneidad con X^2 y test G.

#########################################################
# Sección modificable por el usuario
#########################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro VII.2.V.csv",encoding="latin1")

# Selección de las variables de interés
varInteres<-c("Sexo","Presencia.del.virus")

# Selección de las variables de agrupación.
varAgrupa<-NULL

# Tipos de pruebas a realizar
# 1. Prueba chi-cuadrado.
# 2. Prueba G sin corrección de Williams.
# 3. Prueba G con corrección de Williams.
# 4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2).
# 5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral.
# 6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor que.
# 7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor que.

pruebas<-c(1,2,4,5,6,7)

# Nombre del archivo de salida con los resultados.
nomSalida<-"Salida Cuadro VII.2.V.txt"
```

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro VII.2.V.csv**.

datos<-read.csv2("Cuadro VII.2.V.csv",encoding="latin1")

CUADRO VII.2. *(Continuación)*[R]

Paso 2. Se seleccionan las dos variables que se quieren estudiar, en este caso el sexo y la presencia o ausencia del virus.

```
varInteres<-c("Sexo","Presencia.del.virus")
```

Paso 3. Es posible hacer los cálculos estadísticos agrupando las variables de la forma que se quiera. En el caso de seleccionar «NULL», no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna, como es el caso de este ejemplo.

```
varAgrupa<-NULL
```

Paso 4. A continuación es necesario especificar los tipos de contrastes que se quieren realizar. En nuestro ejemplo, todos los contrastes a excepción de la prueba G con corrección de Williams.

```
# 1. Prueba chi-cuadrado.
# 2. Prueba G sin corrección de Williams
# 3. Prueba G con corrección de Williams
# 4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2)
# 5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral
# 6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor
que.
# 7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor
que.

pruebas<-c(1,2,4,5,6,7)
```

Paso 5. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
nomSalida<-"Salida Cuadro VII.2.V.txt"
```

Paso 6. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*. Los resultados se muestran en la ventana de la siguiente página.

Chi-cuadrado.
El valor χ^2 es 0,15 con p = 0,694. Como p > 0,05 concluimos que no existe relación entre el sexo y la presencia del virus del herpes labial.

Corrección de Yates
En este caso dado, como las frecuencias de cada categoría son mayores de 5, no sería necesaria esta corrección. Suponiendo que lo fuese, tendríamos un valor de contraste de 0,035 y p = 0,851, por lo que se acepta la hipótesis de que la presencia del virus es igual en ambos sexos.

CUADRO VII.2. *(Continuación)*[R]

Prueba exacta de Fisher

Nos devuelve el estadístico calculado tanto con una hipótesis alternativa unilateral (que informa de si es mayor el número de mujeres portadoras del virus que el número de hombres portadores) como con una hipótesis alternativa bilateral (que simplemente informa si la presencia del virus es distinta). Las probabilidades serían p = 0,425 unilateral y p = 0,838 para bilateral. En ambos casos se acepta la hipótesis de homogeneidad de variables, aunque con nuestros datos (no suponemos a priori si serán más mujeres con virus que hombres), debemos emplear la probabilidad bilateral.

```
$tabla
        Presencia.del.virus
Sexo     No portador Portador
  Hombre          25       34
  Mujer           19       22

$listaPruebas
$listaPruebas[[1]]

        Pearson's Chi-squared test

data:  tbl1
X-squared = 0.1546, df = 1, p-value = 0.6942


$listaPruebas[[2]]

        Log likelihood ratio (G-test) test of independence without correction

data:  tbl1
Log likelihood ratio statistic (G) = 0.1545, X-squared df = 1, p-value
= 0.6943


$listaPruebas[[3]]

        Pearson's Chi-squared test with Yates' continuity correction

data:  tbl1
X-squared = 0.0355, df = 1, p-value = 0.8505


$listaPruebas[[4]]

        Fisher's Exact Test for Count Data

data:  tbl1
p-value = 0.8379
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3541774 2.0532746
sample estimates:
odds ratio
 0.8527715


$listaPruebas[[5]]

        Fisher's Exact Test for Count Data

data:  tbl1
p-value = 0.4248
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 1.804971
sample estimates:
odds ratio
 0.8527715


$listaPruebas[[6]]

        Fisher's Exact Test for Count Data

data:  tbl1
p-value = 0.7253
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.4028887       Inf
sample estimates:
```

CUADRO VII.2. *(Continuación)*[STATISTICA]

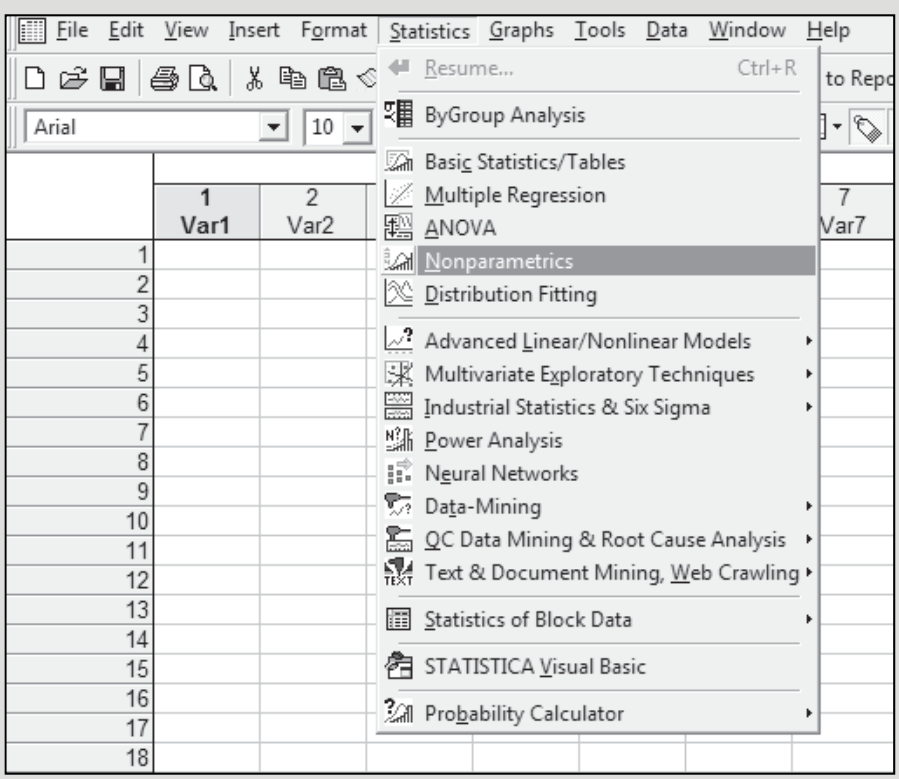
Aplicación con STATISTICA.

EJEMPLO. Se van a utilizar los mismos datos que se usaron con R, es decir, determinar si existen diferencias en la frecuencia de portadores del virus del herpes labial entre hombres y mujeres.

Paso 1. En el caso de STATISTICA, en tablas 2x2, es mejor preparar la tabla de contingencia como se muestra a continuación, para realizar el análisis de forma más fácil.

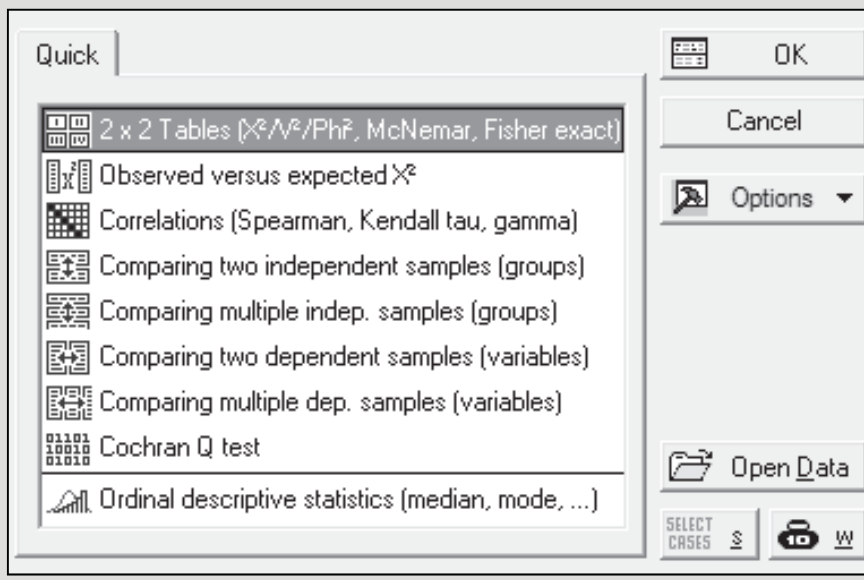
	Portador del virus		
	NO	**SI**	**Total**
Hombres	25	34	**59**
Mujeres	19	22	**45**
Total	**44**	**60**	**104**

Paso 2. En el programa STATISTICA no es necesario introducir datos en la hoja de trabajo para analizar tablas de contingencia 2x2. Se selecciona «Statistics» ⇨ «Nonparametrics».

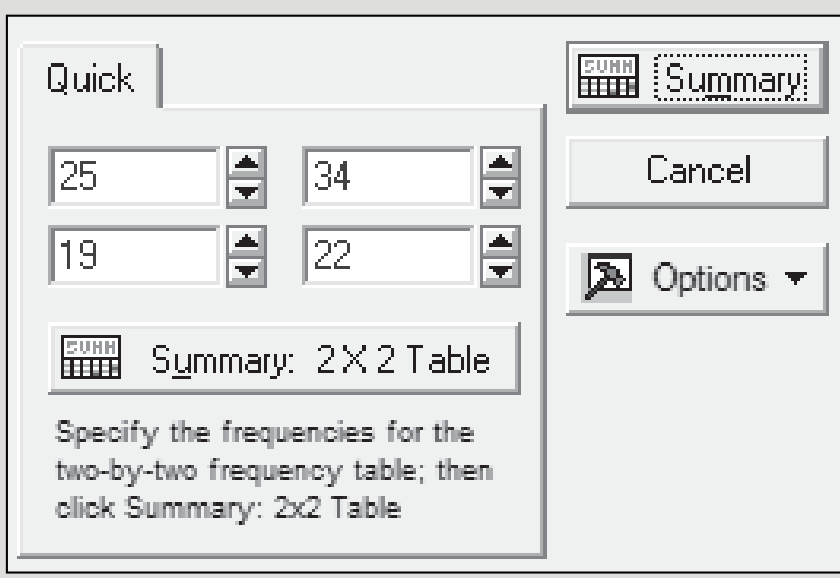


CUADRO VII.2. *(Continuación)* STATISTICA

Paso 3. En la ventana que aparece seleccionamos «2x2 Tables (X^2/V^2/Phi^2, McNemar, Fisher exact».



Paso 4. En las casillas se introducen los valores de la tabla de contingencia.



CUADRO VII.2. *(Continuación)*[STATISTICA]

Paso 5. Pulsando sobre el icono "Summary 2X2 Table" podemos ver los resultados.

	2 x 2 Table (Spreadsheet9)		
	Column 1	Column 2	Row Totals
Frequencies, row 1	25	34	59
Percent of total	25,000%	34,000%	59,000%
Frequencies, row 2	19	22	41
Percent of total	19,000%	22,000%	41,000%
Column totals	44	56	100
Percent of total	44,000%	56,000%	
Chi-square (df=1)	,15	p= ,6942	
V-square (df=1)	,15	p= ,6956	
Yates corrected Chi-square	,04	p= ,8506	
Phi-square	,00155		
Fisher exact p, one-tailed		p= ,4248	
two-tailed		p= ,8379	
McNemar Chi-square (A/D)	,09	p= ,7705	
Chi-square (B/C)	3,70	p= ,0545	

Chi-cuadrado.
El valor χ^2 es 0,15 con p = 0,694. Como p > 0,05 concluimos que no existe relación entre si es hombre o mujer y la presencia del virus del herpes labial.

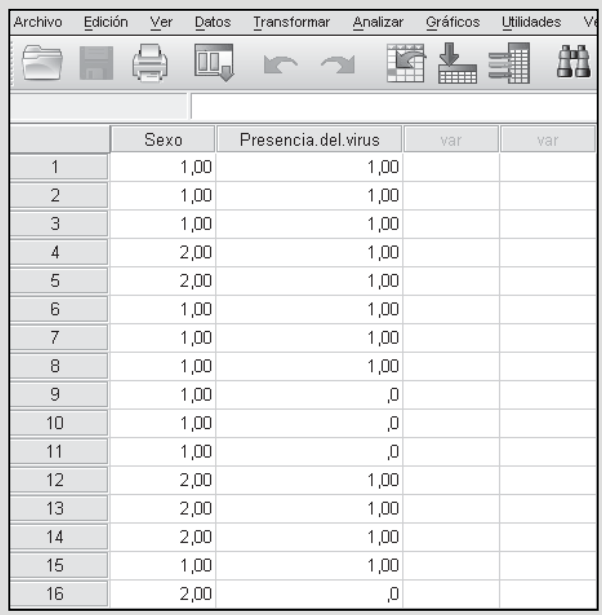
Corrección de Yates
En este caso dado, como las frecuencias de cada categoría son mayores de 5, no sería necesaria esta corrección. Suponiendo que lo fuese, tendríamos un valor de contraste de 0,04 y p = 0,851, por lo que se acepta la hipótesis de que la presencia del virus es igual en ambos sexos.

Prueba exacta de Fisher
Nos devuelve el estadístico calculado tanto con una hipótesis alternativa unidireccional (*one-tailed*, que supone a priori que es mayor la proporción de mujeres portadoras del virus que la de hombres) como con una hipótesis alternativa bilateral – *two tailed* (que simplemente afirma que la presencia del virus es distinta). Las probabilidades serían p = 0,425 unilateral y p = 0,838 para bilateral. En ambos casos se acepta la hipótesis de homogeneidad de variables, aunque con nuestros datos (no suponemos a priori si serán más mujeres con virus que hombres), debemos emplear la probabilidad bilateral.

CUADRO VII.2. *(Continuación)*[SPSS]

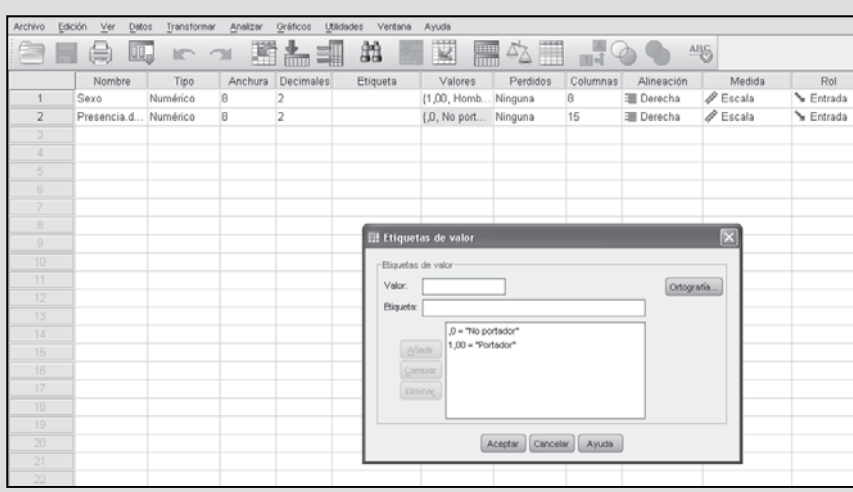
Aplicación con SPPS.

EJEMPLO. Vamos a utilizar los mismos datos que en el caso de la estimación con R y con STATISTICA, los cuales se encuentran en el archivo **Cuadro VII.2.sav** y se muestran en la siguiente ventana.



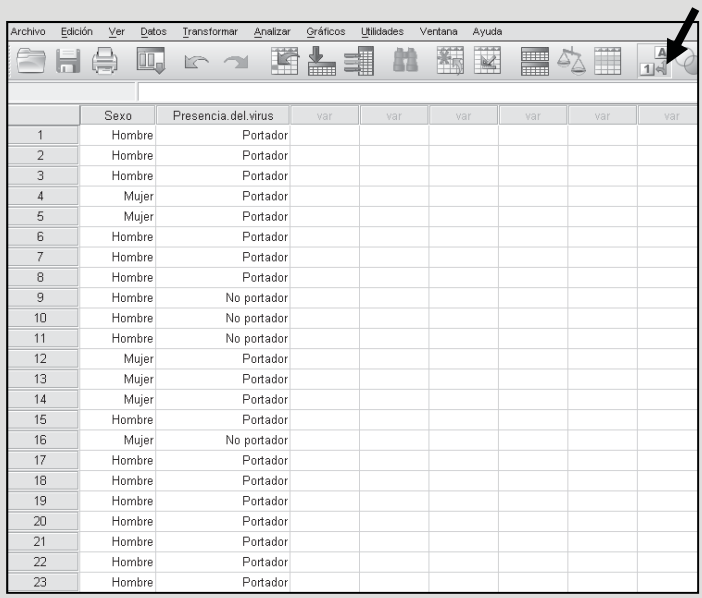
	Sexo	Presencia.del.virus	var	var
1	1,00	1,00		
2	1,00	1,00		
3	1,00	1,00		
4	2,00	1,00		
5	2,00	1,00		
6	1,00	1,00		
7	1,00	1,00		
8	1,00	1,00		
9	1,00	,0		
10	1,00	,0		
11	1,00	,0		
12	2,00	1,00		
13	2,00	1,00		
14	2,00	1,00		
15	1,00	1,00		
16	2,00	,0		

Paso 1. Al seleccionar la pestaña «Vista de variables» aparece la siguiente ventana donde es posible especificar el nombre de cada variable en «Nombre», el tipo de variable en «Tipo», etc. En «Valores» aparece la ventana que se muestra a continuación, donde es posible etiquetar a que corresponde cada valor en cada una de las variables.


Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana Ayuda
Nombre Tipo Anchura Decimales Etiqueta Valores Perdidos Columnas Alineación Medida Rol
1 Sexo Numérico 8 2 {1,00, Homb... Ninguna 8 Derecha Escala Entrada
2 Presencia.d... Numérico 8 2 {,0, No port... Ninguna 15 Derecha Escala Entrada
Etiquetas de valor
Etiquetas de valor
Valor:
Etiqueta:
Ortografía...
,0 = "No portador"
1,00 = "Portador"
Añadir
Cambiar
Eliminar
Aceptar
Cancelar
Ayuda

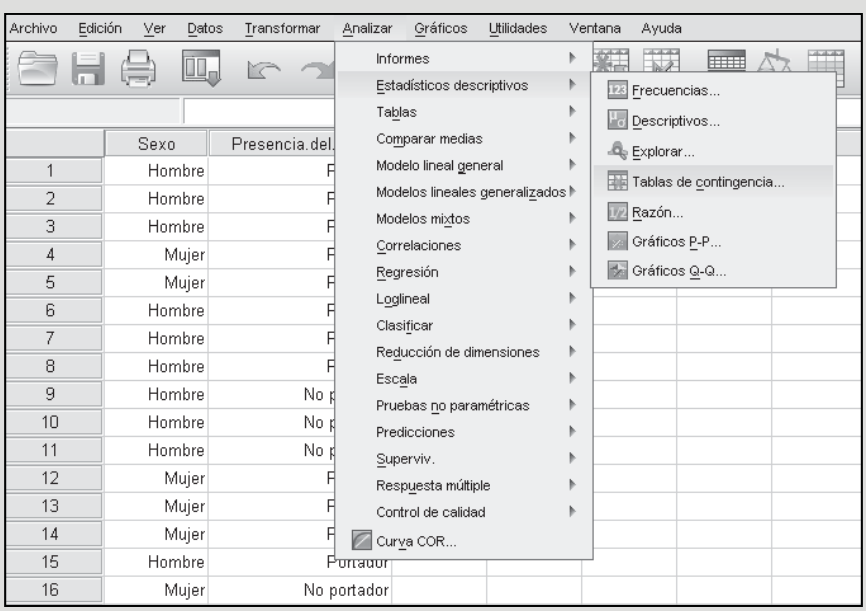
CUADRO VII.2. *(Continuación)*[SPSS]

Paso 2. Al volver a la ventana de la pestaña «Vista de datos», si pulsamos donde indica la flecha, aparecen en vez de los números las etiquetas que corresponden a cada valor.



	Sexo	Presencia.del.virus	var	var	var	var	var	var
1	Hombre	Portador						
2	Hombre	Portador						
3	Hombre	Portador						
4	Mujer	Portador						
5	Mujer	Portador						
6	Hombre	Portador						
7	Hombre	Portador						
8	Hombre	Portador						
9	Hombre	No portador						
10	Hombre	No portador						
11	Hombre	No portador						
12	Mujer	Portador						
13	Mujer	Portador						
14	Mujer	Portador						
15	Hombre	Portador						
16	Mujer	No portador						
17	Hombre	Portador						
18	Hombre	Portador						
19	Hombre	Portador						
20	Hombre	Portador						
21	Hombre	Portador						
22	Hombre	Portador						
23	Hombre	Portador						

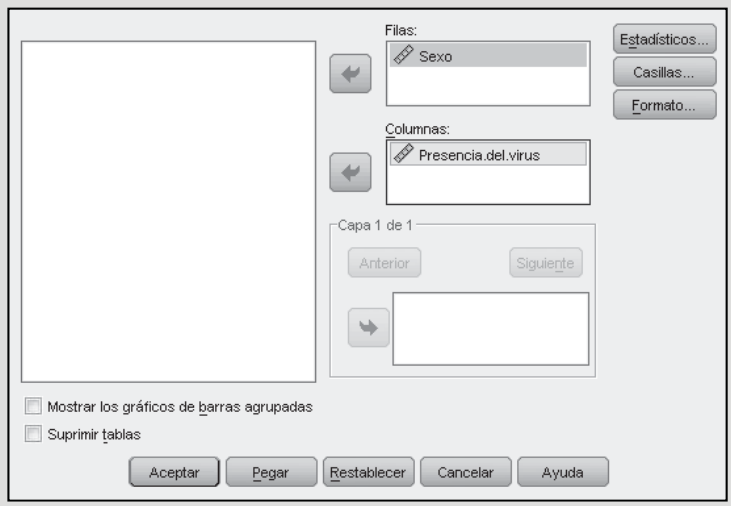
Paso 3. A continuación se selecciona «Analizar» ⇨ «Estadísticos descriptivos» ⇨ «Tablas de contingencia…».



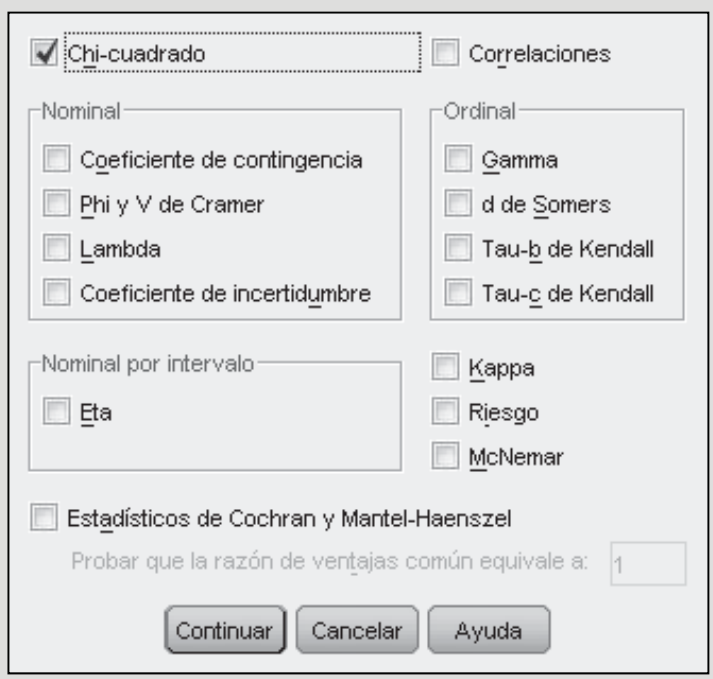
	Sexo	Presencia.del.
1	Hombre	P
2	Hombre	P
3	Hombre	P
4	Mujer	P
5	Mujer	P
6	Hombre	P
7	Hombre	P
8	Hombre	P
9	Hombre	No p
10	Hombre	No p
11	Hombre	No p
12	Mujer	P
13	Mujer	P
14	Mujer	P
15	Hombre	Portador
16	Mujer	No portador

CUADRO VII.2. *(Continuación)*[SPSS]

Paso 4. En «Filas» hay que poner la variable que tiene los grupos a comparar, en este ejemplo «Sexo». En «Columnas» la variable que se quiere comparar, en este caso «Presencia.del.virus».



Paso 5. En la ventana anterior hay que pulsar en «Estadísticos…» y seleccionar «Chi-cuadrado».



CUADRO VII.2. *(Continuación)*[SPSS]

Paso 6. En la tabla de contingencia aparecen las frecuencias de todas las clases.

Tabla de contingencia Sexo * Presencia.del.virus

Recuento

		Presencia.del.virus		
		No portador	Portador	Total
Sexo	Hombre	25	34	59
	Mujer	19	22	41
Total		44	56	100

Paso 7. Los resultados de las pruebas se muestran en la siguiente tabla.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	,155[a]	1	,694		
Corrección por continuidad[b]	,036	1	,851		
Razón de verosimilitudes	,154	1	,694		
Estadístico exacto de Fisher				,838	,425
Asociación lineal por lineal	,153	1	,696		
N de casos válidos	100				

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 18,04.

b. Calculado sólo para una tabla de 2x2.

Chi-cuadrado. El valor χ^2 es 0,155 con p = 0,694. Como $p > 0,05$ concluimos que no existen diferencias significativas en el número de portadores del virus entre hombres y mujeres.

Corrección de Yates. En SPSS se denomina "corrección por continuidad". En este ejemplo, como las frecuencias de cada categoría son mayores que 5, no sería necesaria esta corrección. Suponiendo que lo fuese, tendríamos un valor de contraste de 0,036 y p = 0,851, por lo que se acepta la hipótesis nula de que la presencia del virus es igual en ambos sexos.

Prueba exacta de Fisher. Nos devuelve el estadístico calculado tanto con una hipótesis alternativa unidireccional (*one-tailed*, que supone a priori que es mayor la proporción de mujeres portadoras del virus que la de hombres) como con una hipótesis alternativa bilateral – *two tailed* (que simplemente afirma que la presencia del virus es distinta). Las probabilidades serían p = 0,425 unilateral y p = 0,838 para bilateral. En ambos casos se acepta la hipótesis de homogeneidad de variables, aunque con nuestros datos (no suponemos a priori si serán más mujeres con virus que hombres), debemos emplear la probabilidad bilateral.

VII.2. Muestras relacionadas

Se puede dar el caso de que las variables a comparar estén relacionadas, es decir, son datos pareados. Por ejemplo, cuando a un grupo de personas se les aplica un tratamiento y se estudia si ese fármaco provoca algún tipo de alergia, cuantificada esta alergia por medio de alguna variable cualitativa como presencia o ausencia de manchas en la piel. En este caso, las variables a comparar no son independientes, ya que son las mismas personas las que se comparan a lo largo del tiempo.

Al igual que en el caso de las muestras independientes, existen pruebas para variables politómicas y dicotómicas.

VII.2.1. VARIABLES POLITÓMICAS

VII.2.1.1. ANOVA de Friedman

El test desarrollado por Friedman (1937) es la prueba análoga al análisis de varianza (ANOVA) para muestras repetidas por lo que, a pesar de que se trata de un contraste χ^2, recibe el nombre de ANOVA.

Los pasos de este estadístico están descritos por Siegel & Castellán (1988). De forma resumida son:

1. Asignar rangos dentro de cada elemento de la muestra.
2. Sumar los rangos de cada elemento de la muestra.
3. Calcular el estadístico χ^2 de contraste y compararlo con el valor tabulado χ^2 crítico con *a* - 1 grados de libertad, siendo *a* el número de variables.

El método para aplicar este test se explica en el Cuadro VII.3.

CUADRO VII.3. ANOVA de Friedman

Aplicación con R.

EJEMPLO. En el archivo **Cuadro VII.3.V.csv** hay datos sobre como evalúan 4 diferentes tipos de vinos 15 enólogos. La categoría 1 es "Muy malo", 2 es "Malo", 3 es "Bueno" y 4 "Muy bueno". El objetivo es determinar si existen diferencias entre la calidad de los vinos. Como se compara la evaluación de cada uno de los enólogos de los distintos vinos, los datos no son independientes, están relacionados. Además cada una de las variables tiene más de 2 valores, por tanto son politómicas.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro VII.3.R**, que se muestra en la siguiente ventana. Solo es necesario introducir los datos que se solicitan en la sección modificable por el usuario.

CUADRO VII.3. *(Continuación)*[R]

```
# Cuadro VII.3
# Prueba de Friedman

#########################################################
# Sección modificable por el usuario
#########################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro VII.3.V.csv",encoding="latin1")

# Selección de las variables de interés
# Si se quiere realizar la prueba de Friedman
# Las variables de interés deben seguir en estricto orden:
# Variable RESPUESTAS, FACTOR, BlOQUE
varInteres<-c("Calidad","Vino","Enólogo")

# Selección de las variables de agrupación.
varAgrupa<-NULL

# Tipos de pruebas a realizar
# 1. Prueba chi-cuadrado.
# 2. Prueba G sin corrección de Williams.
# 3. Prueba G con corrección de Williams.
# 4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2).
# 5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral.
# 6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor que.
# 7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor que.
# 8. Prueba ANOVA de Friedman.

pruebas<-c(8)

# Nombre del archivo de salida con los resultados.
nomSalida<-"Salida Cuadro VII.3.V.txt"
```

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro VII.3.V.csv**.

datos<-read.csv2("Cuadro VII.3.V.csv",encoding="latin1")

Paso 2. Se seleccionan las dos variables que se quieren estudiar. En el caso del ANOVA de Friedman es necesario ponerlas en el siguiente orden: la variable que tiene los datos de la respuesta, el factor y, por último, la variable que tiene la información sobre los individuos a los que pertenece cada respuesta.

varInteres<-c("Calidad","Vino","Enólogo")

Paso 3. Es posible hacer los cálculos estadísticos agrupando las variables de la forma que se quiera. En el caso de seleccionar «NULL», no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna, como es el caso de este ejemplo.

varAgrupa<-NULL

CUADRO VII.3. *(Continuación)*[R]

Paso 4. A continuación es necesario especificar los tipos de contrastes que se quieren realizar. En nuestro ejemplo, solo se selecciona el ANOVA de Friedman.

```
# 1. Prueba chi-cuadrado.
# 2. Prueba G sin corrección de Williams
# 3. Prueba G con corrección de Williams
# 4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2)
# 5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral
# 6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor que.
# 7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor que.
# 8. Prueba ANOVA de Friedman

pruebas<-c(8)
```

Paso 5. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
nomSalida<-"Salida Cuadro VII.3.V.txt"
```

Paso 6. Los resultados se muestran en la siguiente ventana.

```
$tabla
          medias  desv.est
Vino 1 3.0666667 0.9611501
Vino 2 1.7333333 0.7037316
Vino 3 2.2000000 0.8618916
Vino 4 2.8666667 1.3020131

$listaPruebas
$listaPruebas[[1]]

        Friedman rank sum test

data:  varInt, factor(varFact) and factor(varBloque)
Friedman chi-squared = 10.1172, df = 3, p-value = 0.01760
```

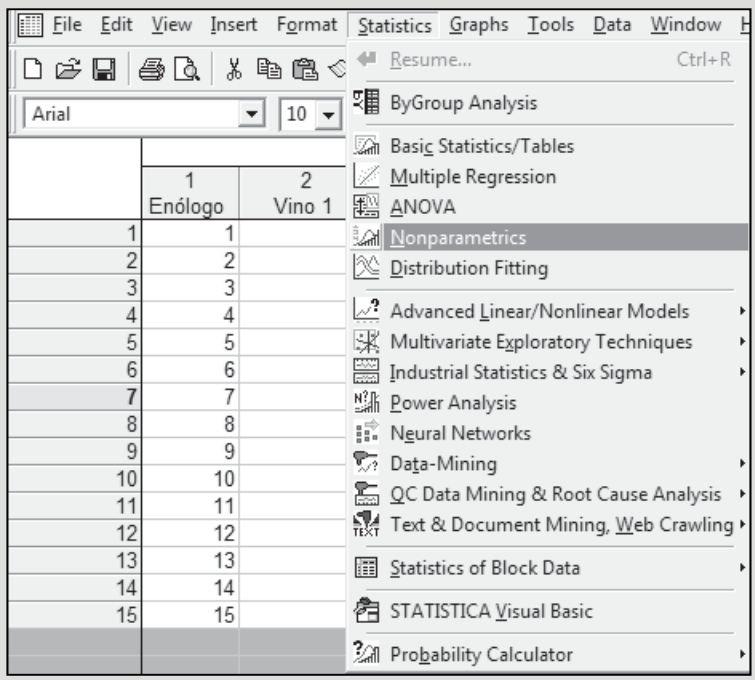
Como la probabilidad es menor de 0,05, se rechaza la hipótesis nula de homogeneidad y, por tanto, hay diferencias significativas entre la calidad de los vinos según los enólogos (ANOVA χ^2 =10,1, *df* = 3, *p* = 0,018).

CUADRO VII.3. *(Continuación)*[STATISTICA]

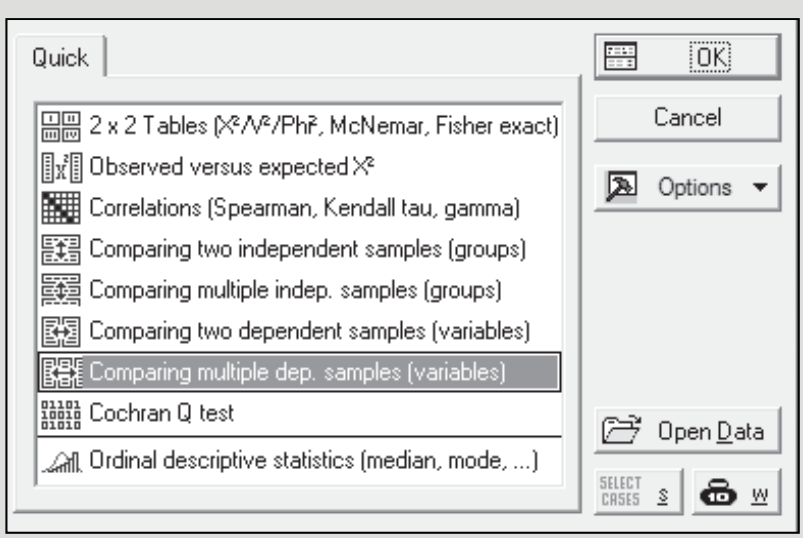
Aplicación con STATISTICA.

EJEMPLO. Se va a utilizar los mismos datos que se usaron con R, es decir, si la calidad de 4 vinos es diferente en función de la valoración dada por 15 enólogos. Los datos están en el archivo **Cuadro VII.3.sta**.

Paso 1. Se selecciona «Statistics» ⇨ «Nonparametrics».

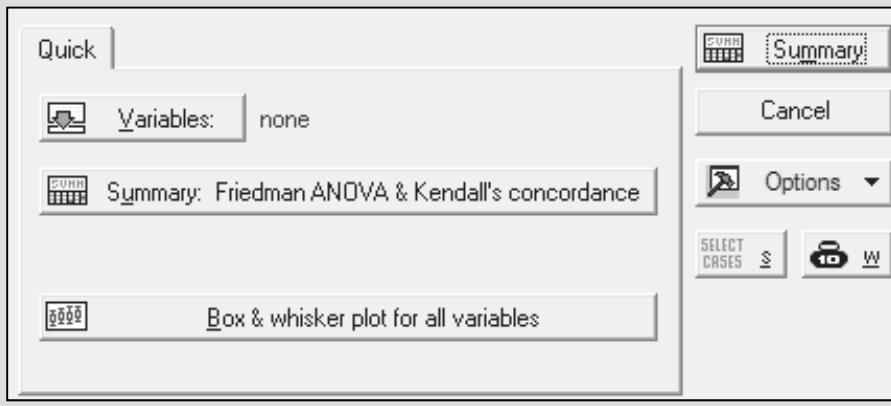


Paso 2. A continuación «Comparing multiple dep. samples (variables)».

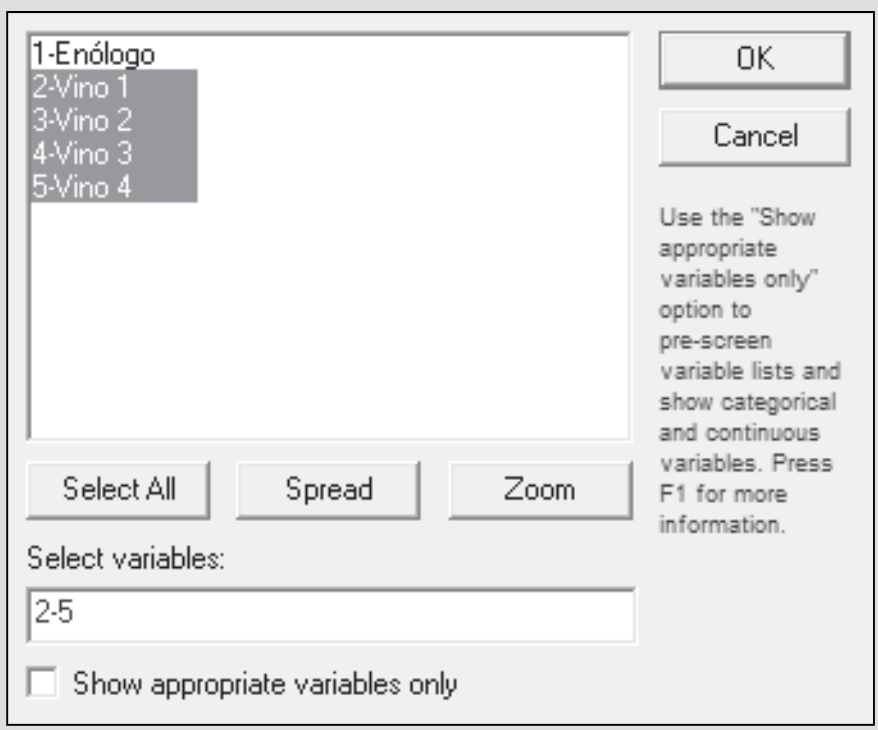


CUADRO VII.3. *(Continuación)*[STATISTICA]

Paso 3. En la ventana que aparece hay que seleccionar «Variables».



Paso 4. Se seleccionan los 4 tipos de vinos y en la pantalla anterior pulsamos el icono «Summary: Friedman ANOVA & Kendall´s concordance».



CUADRO VII.3. *(Continuación)*[SPSS]

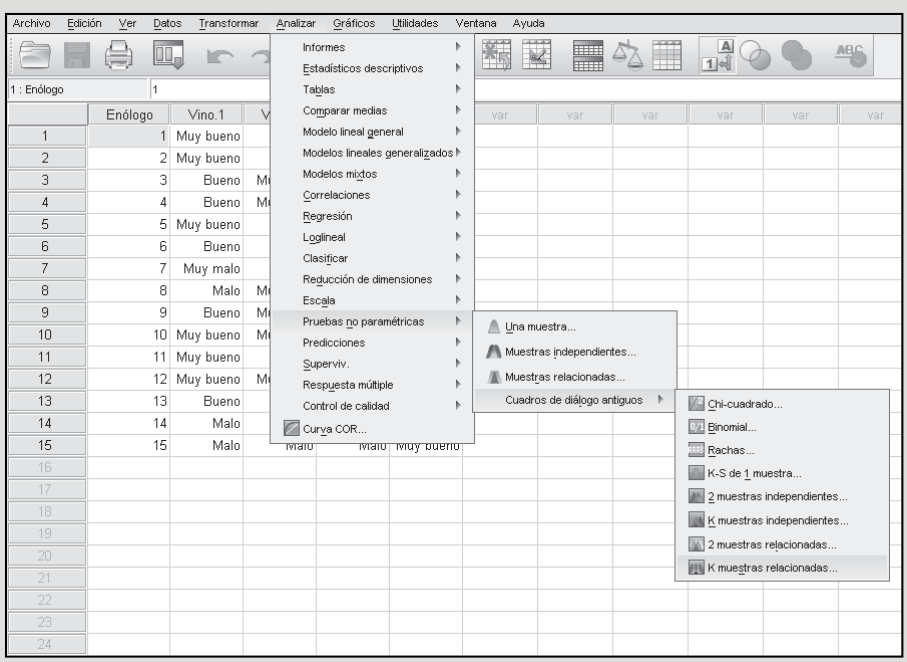
Paso 5. Como la probabilidad es menor de 0,05, se rechaza la hipótesis nula de homogeneidad y, por tanto, hay diferencias significativas entre la calidad de los vinos según los enólogos (ANOVA χ^2 =10,1, *df* = 3, *p* = 0,018).

Variable	Friedman ANOVA and Kendall Coeff. of Concordance (Cuadro VII.3.sta) ANOVA Chi Sqr. (N = 15, df = 3) = 10,11724 p = ,01760 Coeff. of Concordance = ,22483 Aver. rank r = ,16946			
	Average Rank	Sum of Ranks	Mean	Std.Dev.
Vino 1	3,100000	46,50000	3,066667	0,961150
Vino 2	1,766667	26,50000	1,733333	0,703732
Vino 3	2,266667	34,00000	2,200000	0,861892
Vino 4	2,866667	43,00000	2,866667	1,302013

Aplicación con SPPS.

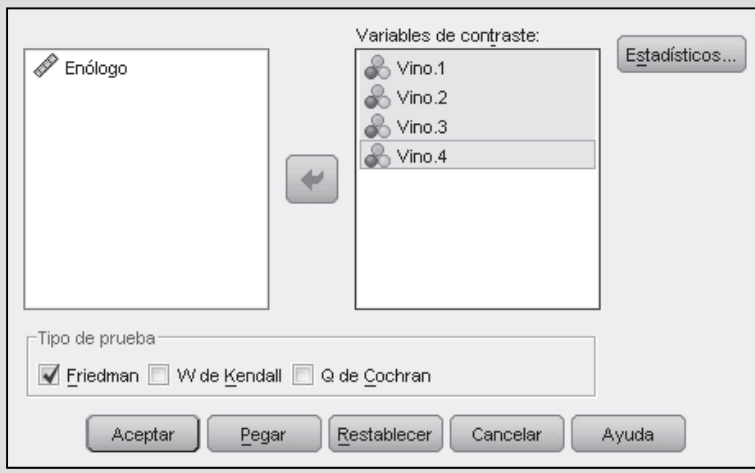
EJEMPLO. Vamos a utilizar los mismos datos que en el caso de la estimación con R y con STATISTICA, los cuales se encuentran en el archivo **Cuadro VII.3.sav**.

Paso 1. Se selecciona «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «K muestras relacionadas…».



CUADRO VII.3. *(Continuación)*[SPSS]

Paso 2. En «Variables de contraste:» se incluyen las variables que se van a analizar y en «Tipo de prueba» se selecciona «Friedman».



Paso 3. Los resultados muestran que la probabilidad es menor que 0,05, por tanto hay diferencias significativas entre la calidad de los vinos (ANOVA χ^2 =10,1, *df* = 3, *p* = 0,018).

Estadísticos de contraste[a]

N	15
Chi-cuadrado	10,117
gl	3
Sig. asintót.	,018

a. Prueba de Friedman

VII.2.2. VARIABLES DICOTÓMICAS

VII.2.2.1. Prueba de McNemar

Se utiliza con muestras a las que se les mide la misma variable dicotómica dos veces, es decir, son datos pareados (McNemar, 1947). Es muy utilizado en medicina para conocer si un tratamiento tiene efecto, pudiéndose representar los datos en una tabla de contingencia 2x2 (véase Tabla VII.2).

Se realiza el siguiente contraste, en el que se asume como hipótesis nula que las frecuencias esperadas *b* y *c* son iguales:

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

El valor crítico de contraste se corresponde con la distribución χ^2 para 1 grado de libertad. Su cálculo se muestra en el Cuadro VII.4.

CUADRO VII.4. Prueba de McNemar

Aplicación con R.

EJEMPLO. El objetivo del estudio es determinar si un antibiótico provoca una respuesta alérgica en la piel. Para ello se administró el antibiótico a 50 personas y a otras 50 personas se les dio un placebo. Los resultados comparan la respuesta positiva (alergia) o negativa (no alergia) en las 100 personas entre los días 5 y 10 después del tratamiento. Lo que se contrasta es si hay diferencias en cada uno de los grupos, es decir si son homogéneos en cuanto a la respuesta alérgica, en los dos períodos considerados, 5 y 10 días. Si suponemos que 5 días es "antes" y 10 días "después" y la alergia aparece necesariamente en ese intervalo, el efecto alérgico del antibiótico puede probarse si encontramos homogeneidad en el grupo placebo y no homogeneidad en el grupo al que se le suministró antibiótico, aunque en este último caso debemos comprobar adicionalmente que la alergia es mayor "después" que "antes" y no al revés (en ambos casos el contraste puede ser significativo). Los datos están en el archivo **Cuadro VII.4.V.csv**.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro VII.4.R**, que se muestra en la siguiente ventana.

```
# Cuadro VII.4
# Prueba de McNemar

#########################################################
# Sección modificable por el usuario
#########################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro VII.4.V.csv",encoding="latin1")

# Selección de las variables de interés
# Si se quiere realizar la prueba de Friedman
# Las variables de interés deben seguir en estricto orden:
# Variable RESPUESTAS, FACTOR, BlOQUE
varInteres<-c("Día.5","Día.10")

# Selección de las variables de agrupación.
varAgrupa<-c("Antibiótico")

# Tipos de pruebas a realizar
#  1. Prueba chi-cuadrado.
#  2. Prueba G sin corrección de Williams.
#  3. Prueba G con corrección de Williams.
#  4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2).
#  5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral.
#  6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor que.
#  7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor que.
#  8. Prueba ANOVA de Friedman.
#  9. Prueba de McNemar sin corrección de continuidad.
# 10. Prueba de McNemar con corrección de continuidad (Sólo para tablas 2x2).

pruebas<-c(9,10)
```

CUADRO VII.4. *(Continuación)*^{R}

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro VII.4.V.csv**.

```
datos<-read.csv2("Cuadro VII.4.V.csv",encoding="latin1")
```

Paso 2. Se seleccionan las dos variables que se quieren estudiar, en este caso se comparan los días 5 y 10 del experimento.

```
varInteres<-c("Día.5","Día.10")
```

Paso 3. Es posible hacer los cálculos estadísticos agrupando las variables de la forma que se quiera. En el caso de seleccionar «NULL», no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna. En nuestro ejemplo se quiere hacer un análisis para cada grupo de tratamiento, es decir, para las personas a las que se les suministró y a las que no se les suministró el antibiótico.

```
varAgrupa<-c("Antibiótico")
```

Paso 4. A continuación es necesario especificar los tipos de contrastes que se quieren realizar. En nuestro ejemplo, las pruebas de McNemar con y sin corrección de continuidad.

```
#  1. Prueba chi-cuadrado.
#  2. Prueba G sin corrección de Williams
#  3. Prueba G con corrección de Williams
#  4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2).
#  5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral
#  6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor
que.
#  7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor
que.
#  8. Prueba ANOVA de Friedman
#  9. Prueba de McNemar sin corrección de continuidad.
# 10. Prueba de McNemar con corrección de continuidad (Sólo para tablas
2x2).

pruebas<-c(9,10)
```

Paso 5. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
nomSalida<-"Salida Cuadro VII.4.V.txt"
```

Paso 6. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*. Los resultados se muestran en la siguiente ventana.

CUADRO VII.4. *(Continuación)*[R]

En el caso de las personas a las que no se les suministró el antibiótico, como p= 0,248 en el contraste con la corrección de continuidad (que es el que debe usarse cuando hay frecuencias menores que 5), se acepta la hipótesis nula de que no existen diferencias significativas en la respuesta alérgica observada en la piel entre el día 5 y el día 10. En el caso de las personas a las que se les suministró el antibiótico como p = 0,133, al ser mayor de 0,05, también se acepta la hipótesis nula de que no existen diferencias significativas en la respuesta alérgica al antibiótico entre el día 5 y día 10.

```
$NO
$NO$tabla
          Día.10
Día.5      Negativa Positiva
  Negativa       43        3
  Positiva        0        4

$NO$listaPruebas
$NO$listaPruebas[[1]]

        McNemar's Chi-squared test

data:  tbl1
McNemar's chi-squared = 3, df = 1, p-value = 0.08326


$NO$listaPruebas[[2]]

        McNemar's Chi-squared test with continuity correction

data:  tbl1
McNemar's chi-squared = 1.3333, df = 1, p-value = 0.2482




$SI
$SI$tabla
          Día.10
Día.5      Negativa Positiva
  Negativa       39        4
  Positiva        0        7

$SI$listaPruebas
$SI$listaPruebas[[1]]

        McNemar's Chi-squared test

data:  tbl1
McNemar's chi-squared = 4, df = 1, p-value = 0.0455


$SI$listaPruebas[[2]]

        McNemar's Chi-squared test with continuity correction

data:  tbl1
McNemar's chi-squared = 2.25, df = 1, p-value = 0.1336
```

CUADRO VII.4. *(Continuación)*[STATISTICA]

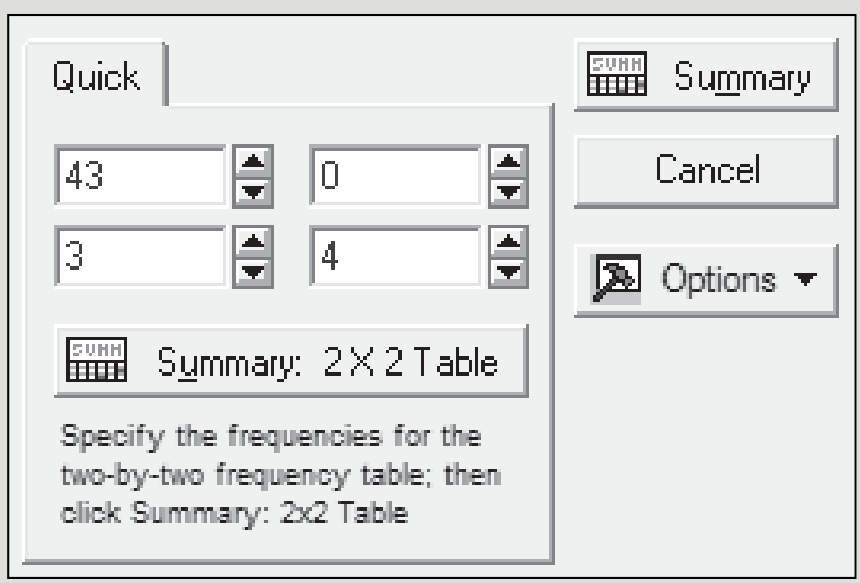
Aplicación con STATISTICA.

EJEMPLO. Se va a utilizar los mismos datos que en la aplicación con R, es decir, determinar si un antibiótico provoca una respuesta alérgica en la piel en un grupo de 100 personas. Los datos están en el archivo **Cuadro VII.4.sta**.

Paso 1. Se puede hacer de dos formas diferentes. Por un lado se puede hacer como se explicó en el Cuadro VII.2 (Pasos 1 a 3) ya que al tratarse de una tabla de contingencia 2x2 se puede introducir la tabla directamente sin necesidad de un archivo de datos. Para ello, como se explicó en el Cuadro VII.2, se selecciona «Statistics» ⇨ «Nonparametrics» y en la ventana que aparece seleccionamos «2x2 Tables (X^2/V^2/Phi^2, McNemar, Fisher exact». Pero como existen dos grupos, las personas a las que no se les suministró y a las que si se les suministró antibiótico, es necesario hacer el proceso dos veces.

La tabla de contingencia para las personas a las que no se les suministró antibiótico y la forma de introducir los datos en STATISTICA se muestran a continuación:

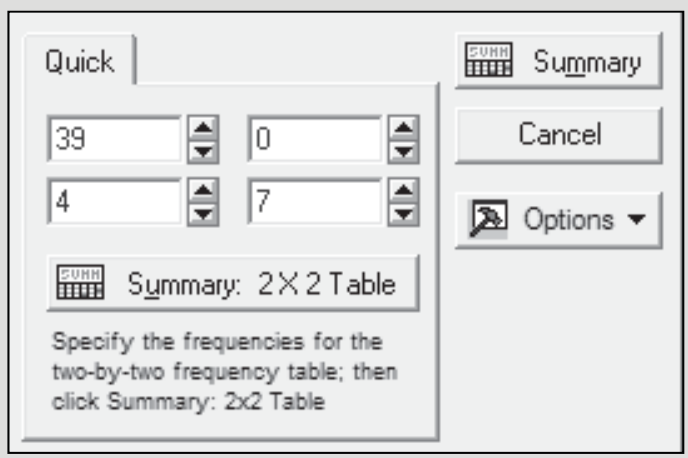
		Día 5		
		-	**+**	**Total**
Día 10	**-**	43	0	**43**
	+	3	4	**7**
Total		**46**	**4**	**50**



CUADRO VII.4. *(Continuación)*[STATISTICA]

La tabla de contingencia para las personas a las que se les suministró antibiótico y como se introducen los datos en STATISTICA se muestran a continuación:

		Día 5		
		-	+	Total
Día 10	**-**	39	0	**39**
	+	4	7	**11**
Total		**43**	**7**	**50**



Paso 2. En cada caso, al pulsar «Summary» aparece la tabla de resultados en la que, del conjunto de estadísticos que calcula STATISTICA, aquí solo mostramos los del test de McNemar.

Para el grupo al que no se le suministró el antibiótico:

	2 x 2 Table (Cuadro VII.4.sta)	
	Column 1	Column 2
McNemar Chi-square (A/D)	30,72	p= ,0000
Chi-square (B/C)	1,33	p= ,2482

Para el grupo al que se le suministró el antibiótico:

	2 x 2 Table (Cuadro VII.4.sta)	
	Column 1	Column 2
McNemar Chi-square (A/D)	20,89	p= ,0000
Chi-square (B/C)	2,25	p= ,1336

CUADRO VII.4. *(Continuación)*[STATISTICA]

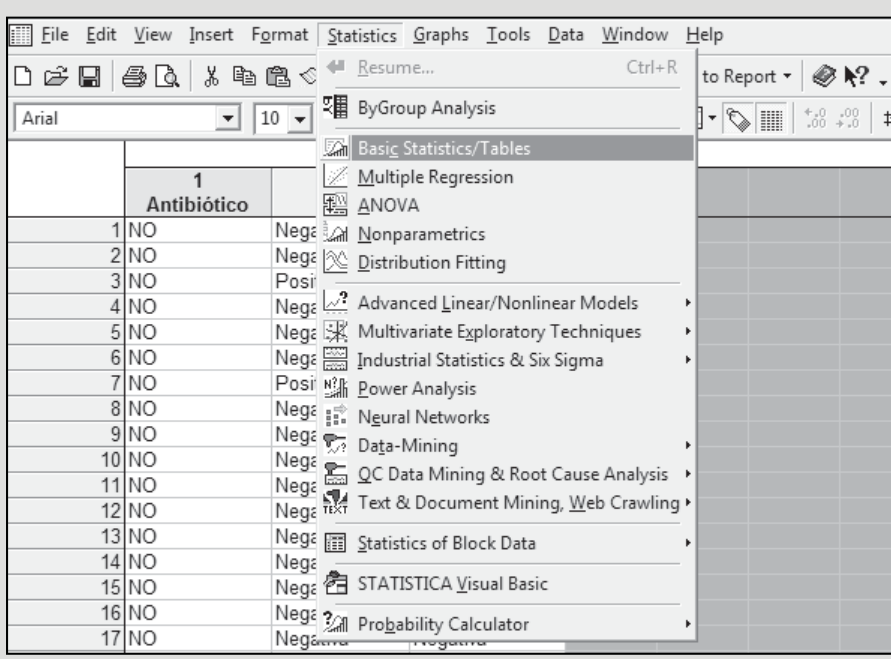
STATISTICA muestra los contrastes con las frecuencias b y c, y también con a y d. Entre los dos contrastes se elige el de las frecuencias b y c, que son las que indican cambios (alergia/no alergia) en el período transcurrido. Si ambas variables están codificadas del mismo modo siempre debe elegirse b y c en lugar de a y d.

En el caso de las personas a las que no se les suministró el antibiótico como $p > 0,05$, se acepta la hipótesis nula de que no existen diferencias significativas en la respuesta alérgica entre el día 5 y el día 10 (McNemar (B/C) = 1,33, $p = 0,248$).

En el caso de las personas a las que se les suministró el antibiótico como también $p > 0,05$, se acepta la hipótesis nula de que no existen diferencias significativas en la respuesta alérgica al antibiótico entre el día 5 y el día 10 (McNemar (B/C) = 2,25, $p = 0,134$).

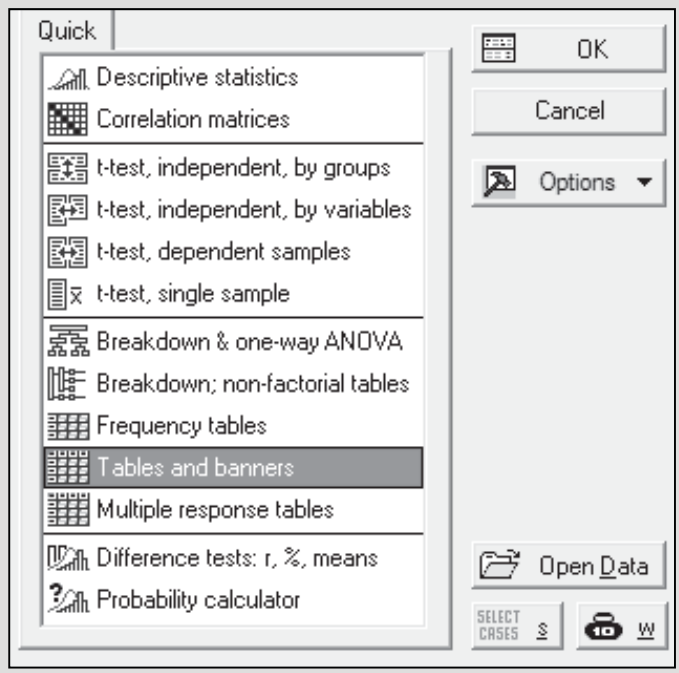
Por tanto, no parece existir una respuesta alérgica al antibiótico, ya que en ese caso deberíamos haber encontrado un efecto significativo en el grupo antibiótico y un efecto no significativo en el grupo placebo.

Paso 3. La otra forma es usando un archivo de datos, más útil en el caso de que existan varios grupos ya que se ahorra el trabajo de tener que hacer varias tablas de contingencia como se mostró anteriormente. Los datos están en el archivo **Cuadro VII.4.sta**. Se selecciona «Statistics» ⇨ «Basic Statistics/Tables».

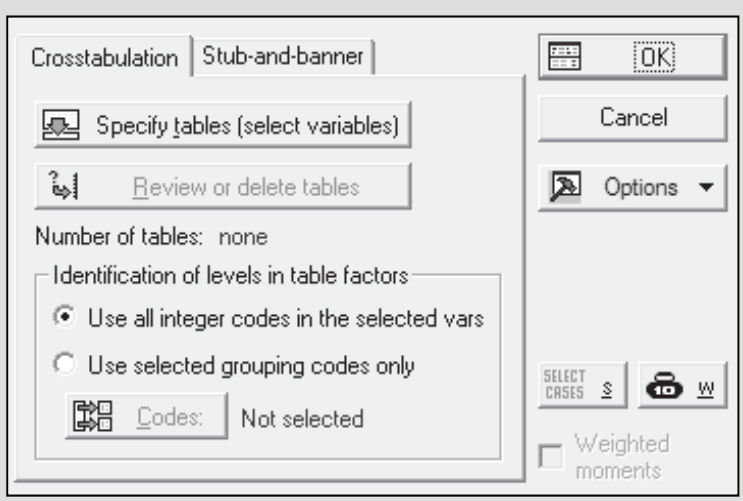


Paso 4. En la ventana que aparece a continuación se selecciona «Tables and banners».

CUADRO VII.4. *(Continuación)* STATISTICA

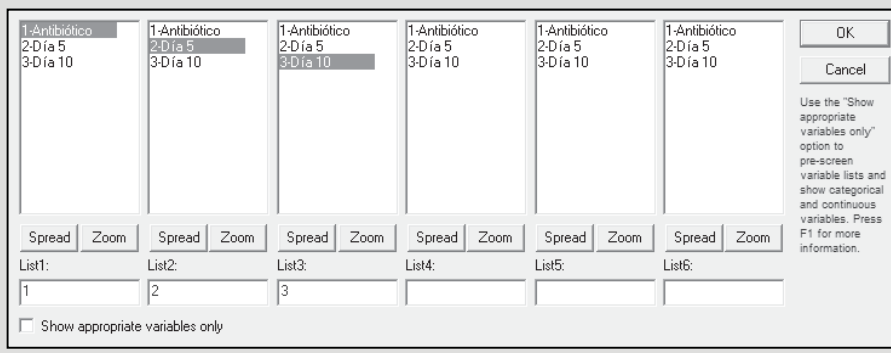


Paso 5. Activamos la pestaña «Crosstabulation» (tabla de contingencia) donde se nos permite seleccionar las categorías con las que queremos trabajar («Use selected grouping codes only») o usar todas las categorías de los datos («Use all integer codes in the selected vars»), que es la opción que trae por defecto el programa. Nosotros trabajaremos con todas las categorías y no modificaremos este punto.

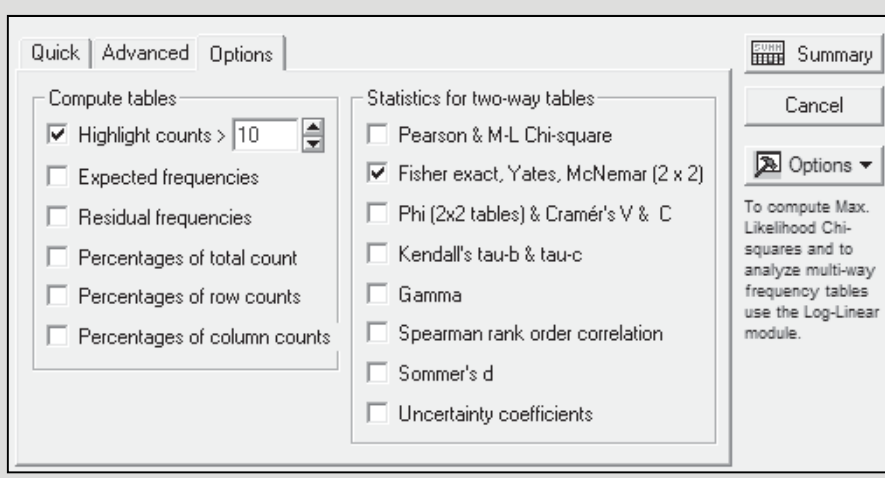


CUADRO VII.4. *(Continuación)*^{STATISTICA}

Paso 6. En la ventana anterior pulsamos sobre «Specify tables (select variables)» y aparece la ventana que se muestra a continuación, donde se pueden seleccionar hasta 6 variables distintas. Como queremos hacer el análisis diferenciando entre antibiótico y placebo, introducimos la variable «Antibiótico» primero. Como lo que se pretende es comparar entre el día 5 y el día 10, estas variables se ponen a continuación. No importa el orden en el que se pongan estas dos variables.

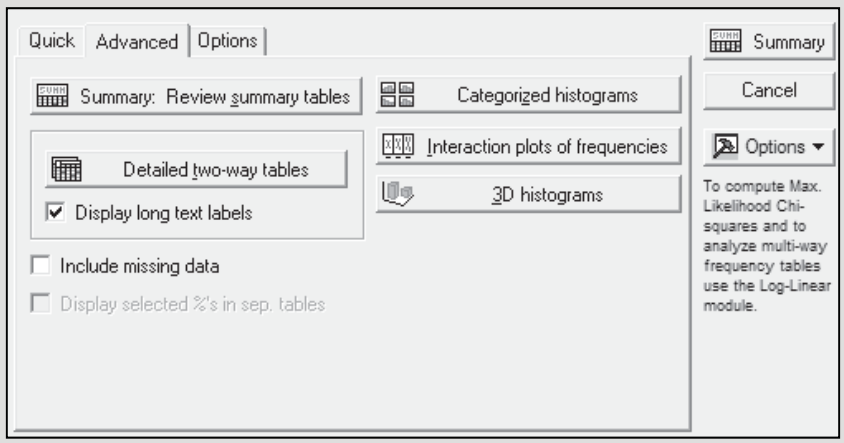


Paso 7. En la ventana del paso 5 pulsamos en OK y nos aparece la ventana que está a continuación, donde hay tres pestañas diferentes, de las cuales dos son de resultados, «Quick» y «Advanced», y una para seleccionar los estadísticos «Options». En la pestaña «Options» seleccionamos la opción «Fisher exact, Yates, McNemar (2 x 2)».



Paso 8. En la pestaña "Advanced" aparece la siguiente ventana, donde pulsando el icono "Summary: Review summary tables" aparece la tabla de contingencia de nuestros resultados.

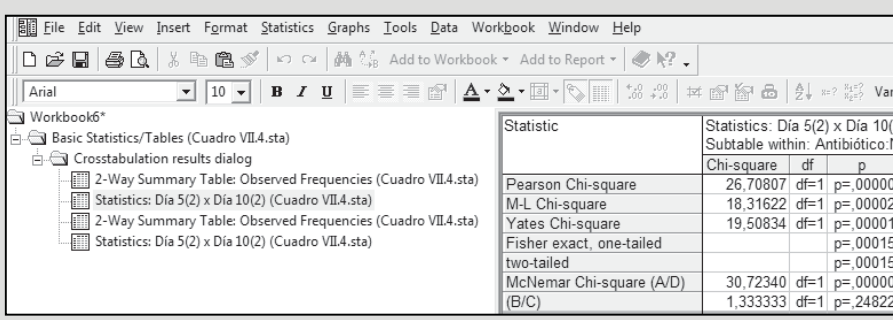
CUADRO VII.4. *(Continuación)*[STATISTICA]



Summary Frequency Table (Cuadro VII.4.sta)
Marked cells have counts > 10
(Marginal summaries are not marked)

Antibiótico	Día 5	Día 10 Negativa	Día 1 Positi
NO	Negativa	43	3
NO	Positiva	0	4
Total		43	7
SI	Negativa	39	4
SI	Positiva	0	7
Total		39	11
Column Total		82	18

Paso 9. En la ventana de la pestaña «Advanced» seleccionamos el icono «Detailed two-ways table» para que el programa nos muestre el resultado de los estadísticos solicitados. Aparece una ventana donde a la izquierda podemos seleccionar los resultados obtenidos para el grupo sin antibiótico y para el grupo con antibiótico. La interpretación de los resultados del test de McNemar se explicó en el paso 2.


File Edit View Insert Format Statistics Graphs Tools Data Workbook Window Help
Add to Workbook
Add to Report
Arial
10
Workbook6*
Basic Statistics/Tables (Cuadro VII.4.sta)
Crosstabulation results dialog
2-Way Summary Table: Observed Frequencies (Cuadro VII.4.sta)
Statistics: Día 5(2) x Día 10(2) (Cuadro VII.4.sta)
2-Way Summary Table: Observed Frequencies (Cuadro VII.4.sta)
Statistics: Día 5(2) x Día 10(2) (Cuadro VII.4.sta)

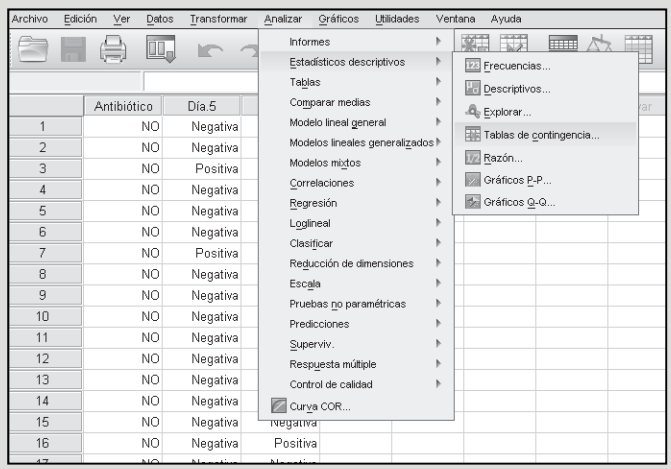
Statistic	Statistics: Día 5(2) x Día 10(2 Subtable within: Antibiótico:N		
	Chi-square	df	p
Pearson Chi-square	26,70807	df=1	p=,00000
M-L Chi-square	18,31622	df=1	p=,00002
Yates Chi-square	19,50834	df=1	p=,00001
Fisher exact, one-tailed			p=,00015
two-tailed			p=,00015
McNemar Chi-square (A/D)	30,72340	df=1	p=,00000
(B/C)	1,333333	df=1	p=,24822

CUADRO VII.4. *(Continuación)*[SPSS]

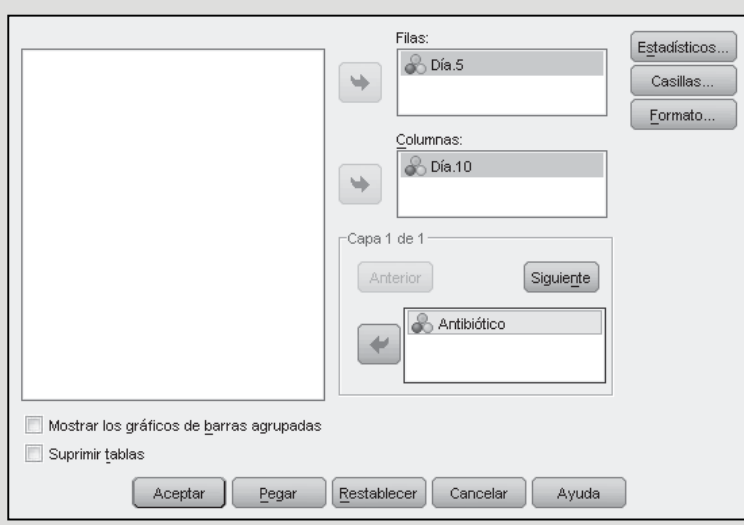
Aplicación con SPSS.

EJEMPLO. Vamos a utilizar los mismos datos que en el caso de la estimación con R y con STATISTICA, los cuales se encuentran en el archivo **Cuadro VII.4.sav**.

Paso 1. Se selecciona «Analizar» ⇨ «Estadísticos descriptivos» ⇨ «Tablas de contingencia…».

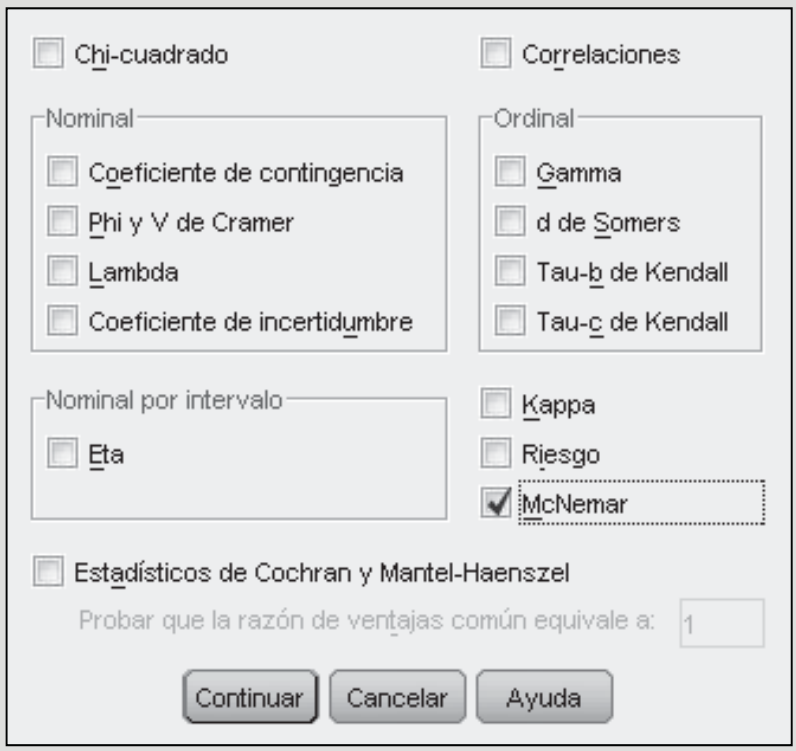


Paso 2. En «Filas» y en «Columnas» hay que poner las variables que se van a contrastar, en este caso «Día.5» y «Día.10», sin importar donde se coloca cada una. Por último, como queremos hacer la distinción entre los grupos con y sin antibiótico, en «Capa» se pone la variable «Antibiótico».



CUADRO VII.4. *(Continuación)*[SPSS]

Paso 3. En la ventana anterior hay que pulsar en «Estadísticos…» y seleccionar «McNemar».



Paso 4. Los resultados se muestran en la siguiente tabla. En el caso de las personas a las que no se les suministró el antibiótico como $p = 0{,}250$, se acepta la hipótesis nula de que no existen diferencias significativas en la respuesta alérgica observada en la piel entre el día 5 y el día 10. En el caso de las personas a las que se les suministró el antibiótico como $p = 0{,}125$, también se acepta la hipótesis nula de que no existen diferencias significativas en la respuesta alérgica al antibiótico entre el día 5 y el día 10.

Pruebas de chi-cuadrado

Antibiótico		Valor	Sig. exacta (bilateral)
NO	Prueba de McNemar		,250[a]
	N de casos válidos	50	
SI	Prueba de McNemar		,125[a]
	N de casos válidos	50	

a. Utilizada la distribución binomial

VII.2.2.2. *Q* de Cochran

Es el equivalente a la prueba de McNemar cuando las mediciones se realizan más de dos veces. Para el cálculo del estadístico de contraste, los datos se presentan en una tabla de contingencia como la que aparece en la Tabla VII.1. Los valores de la tabla de contingencia deben ser transformados a 0 y 1. Por ejemplo, si nuestra variable fuese presencia/ausencia de un carácter, podríamos indicar presencia con 1 y ausencia con 0. El estadístico es el siguiente:

$$Q = \frac{(k-1)\left[k\sum_{j=1}^{k} c_j^2 - \left(\sum_{j=1}^{k} c_j\right)^2\right]}{k\sum_{j=1}^{k} c_j - \sum_{i=1}^{N} n_i^2}$$

Siendo k el número de variables dicotómicas y N el número de casos, c_j el número de éxitos para cada variable y n_i el número de éxitos para cada caso. El método para la obtención de este estadístico aparece en el Cuadro VII.5.

CUADRO VII.5. *Q* de Cochran

Aplicación con R.

EJEMPLO. Continuando con el ejemplo del Cuadro VII.4, que era un estudio para determinar si un antibiótico provoca una respuesta alérgica en la piel, se siguió controlando la respuesta alérgica durante más días a esas 100 personas. Como las mediciones se realizaron más de dos veces, no es posible usar el test de McNemar y se usa el test *Q* de Cochran. Los datos están en el archivo **Cuadro VII.5.V.csv**.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro VII.5.R**, que se muestra en la siguiente ventana.

```
# Cuadro VII.5
# Prueba de Cochran

#########################################################
# Sección modificable por el usuario
#########################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro VII.5.V.csv",encoding="latin1")

# Selección de las variables de interés
# Si se quiere realizar la prueba de Friedman
# Las variables de interés deben seguir en estricto orden:
# Variable RESPUESTAS, FACTOR, BLOQUE
varInteres<-c("Día.5","Día.10","Día.15","Día.20")

# Selección de las variables de agrupación.
varAgrupa<-c("Antibiótico")

# Tipos de pruebas a realizar
#  1. Prueba chi-cuadrado.
#  2. Prueba G sin corrección de Williams.
```

CUADRO VII.5. *(Continuación)*[R]

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro VII.5.V.csv**.

```
datos<-read.csv2("Cuadro VII.5.V.csv",encoding="latin1")
```

Paso 2. Se seleccionan las variables que se quieren estudiar, en este caso se comparan los días 5, 10, 15 y 20 del experimento.

```
varInteres<-c("Día.5","Día.10","Día.15","Día.20")
```

Paso 3. Es posible hacer los cálculos estadísticos agrupando las variables de la forma que se quiera. En el caso de seleccionar «NULL», no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna. En nuestro ejemplo se quiere hacer un análisis para cada grupo de tratamiento, es decir, para las personas a las que se les suministró y a las que no se les suministró el antibiótico.

```
varAgrupa<-c("Antibiótico")
```

Paso 4. A continuación es necesario especificar los tipos de contrastes que se quieren realizar. En nuestro ejemplo, la prueba de Cochran.

```
#  1. Prueba chi-cuadrado.
#  2. Prueba G sin corrección de Williams
#  3. Prueba G con corrección de Williams
#  4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2).
#  5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral
#  6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor
que.
#  7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor
que.
#  8. Prueba ANOVA de Friedman
#  9. Prueba de McNemar sin corrección de continuidad.
# 10. Prueba de McNemar con corrección de continuidad (Sólo para tablas
2x2).
# 11. Prueba de Cochran.

pruebas<-c(11)
```

Paso 5. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
nomSalida<-"Salida Cuadro VII.5.V.txt"
```

Paso 6. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*. La tabla de resultados se muestra a continuación. En el caso de las personas a las que no se les suministró el tratamiento, la probabilidad es menor que 0,05 por lo que rechazamos la hipótesis nula de que las muestras son iguales, es decir, hay diferencias entre los distintos días en los síntomas de alergia detectados en las personas sometidas al placebo (Test Q de Cochran = 9, df = 3, p = 0,03).

CUADRO VII.5. *(Continuación)*[R]

En el caso de las personas a las que se les suministró antibiótico, también se rechaza la hipótesis nula de que las muestras son iguales, es decir, hay diferencias entre los distintos días en los síntomas de alergia detectados (Test *Q* de Cochran = 50,4, *df* = 3, *p* < 0,001).

Es importante resaltar que, el hecho de que se observaran diferencias en el tiempo en ambos grupos (con y sin antibiótico) no significa concluir definitivamente que el antibiótico no es el responsable de la alergia. Esta prueba solo nos permite mostrar que hay diferencias en la respuesta alérgica en el tiempo. Es decir, se podría dar el caso de que en las personas a las que se les suministró placebo, la respuesta alérgica disminuyó en el tiempo y, por el contrario, en las que recibieron antibiótico aumentó. Por tanto, sería necesario a mayores realizar un contraste de independencia como se explica en el Cuadro VIII.1, para poder descartar la asociación entre el antibiótico y la alergia.

```
$NO
$NO$tabla
       Porcentaje Negativa Porcentaje Positiva
Día.5                   92                   8
Día.10                  86                  14
Día.15                  86                  14
Día.20                  84                  16

$NO$listaPruebas
$NO$listaPruebas[[1]]

        Cochran's Q Test for Dependent Samples

data:  datos
Cochran's Q = 9, df = 3, p-value = 0.02929




$SI
$SI$tabla
       Porcentaje Negativa Porcentaje Positiva
Día.5                   86                  14
Día.10                  78                  22
Día.15                  62                  38
Día.20                  38                  62

$SI$listaPruebas
$SI$listaPruebas[[1]]

        Cochran's Q Test for Dependent Samples

data:  datos
Cochran's Q = 50.4, df = 3, p-value = 6.566e-11
```

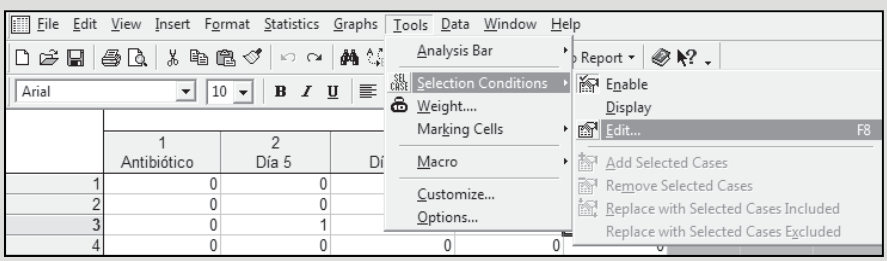
CUADRO VII.5. *(Continuación)*[STATISTICA]

Aplicación con STATISTICA.

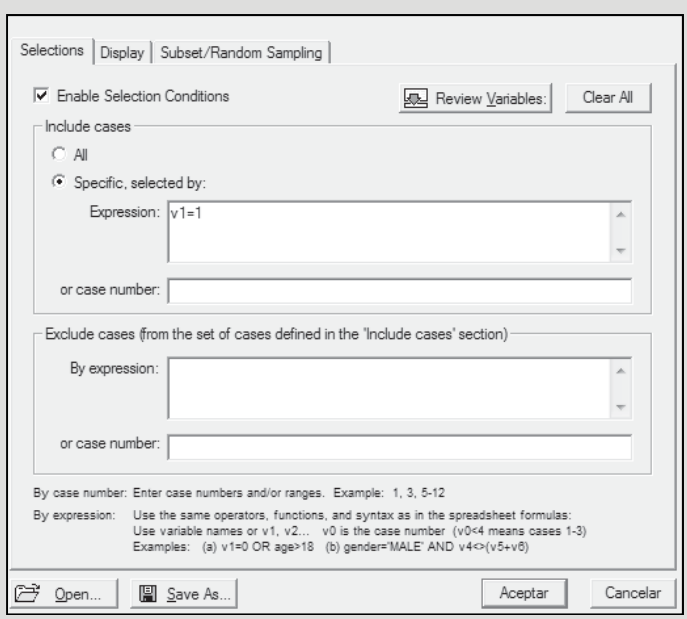
EJEMPLO. Se van a utilizar los mismos datos que en la aplicación con R, es decir, determinar si un antibiótico provoca una respuesta alérgica en la piel en un grupo de 100 personas. En este caso solo vamos a analizar si existen diferencias significativas en la presencia de alergia entre los distintos días en el grupo de personas a las que se les suministró el antibiótico.

Los datos están en el archivo **Cuadro VII.5.sta**.

Paso 1. Debido a que existen dos grupos, las personas a las que se les suministró y a las que no se les suministró el antibiótico, es necesario hacer el análisis por separado. Para seleccionar primero uno de los grupos vamos a «Tools» ⇨ «Selection Conditions» ⇨ «Edit…».

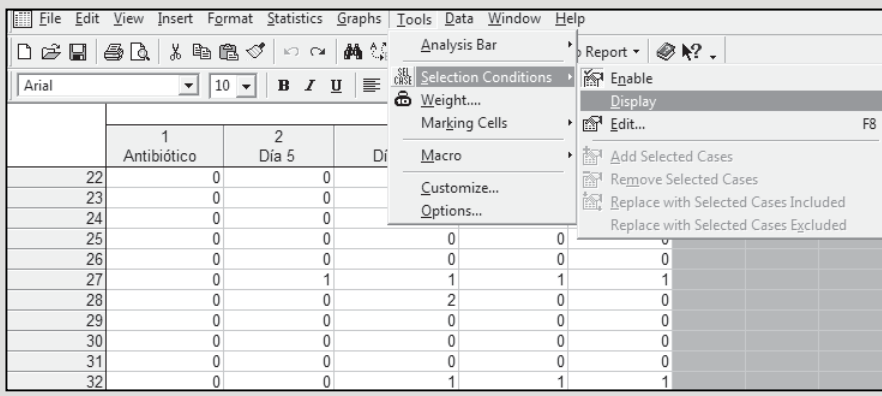


Paso 2. En la pestaña «Selections» se marca «Enable Selection Conditions» ⇨ «Specific, selected by:» donde se selecciona la variable o variables a analizar, usando OR y AND si se quieren combinar varias condiciones. En este caso se selecciona «v1=1», es decir, las personas a las que se les suministró antibiótico.


Selections
Display
Subset/Random Sampling
Enable Selection Conditions
Review Variables:
Clear All
Include cases
All
Specific, selected by:
Expression: v1=1
or case number:
Exclude cases (from the set of cases defined in the 'Include cases' section)
By expression:
or case number:
By case number: Enter case numbers and/or ranges. Example: 1, 3, 5-12
By expression: Use the same operators, functions, and syntax as in the spreadsheet formulas:
Use variable names or v1, v2... v0 is the case number (v0<4 means cases 1-3)
Examples: (a) v1=0 OR age>18 (b) gender='MALE' AND v4<>(v5+v6)
Open...
Save As...
Aceptar
Cancelar

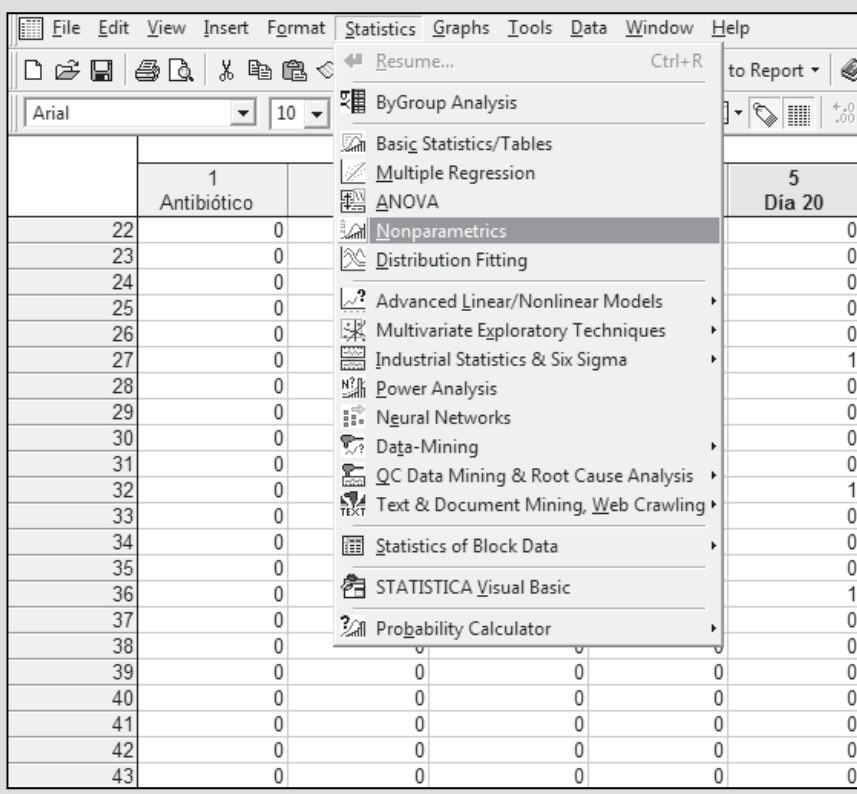
CUADRO VII.5. *(Continuación)* STATISTICA

Paso 3. Para resaltar los casos seleccionados vamos a «Tools» ⇨ «Selection Conditions» ⇨ «Display». En la tabla de datos aparecerán en cursiva los valores seleccionados. En el supuesto de que no aparecieran en cursiva, eso significaría que la selección en el paso 2 no se hizo de forma correcta.



	1 Antibiótico	2 Día 5			
22	0	0			
23	0	0			
24	0	0			
25	0	0	0	0	
26	0	0	0	0	0
27	0	1	1	1	1
28	0	0	2	0	0
29	0	0	0	0	0
30	0	0	0	0	0
31	0	0	0	0	0
32	0	0	1	1	1

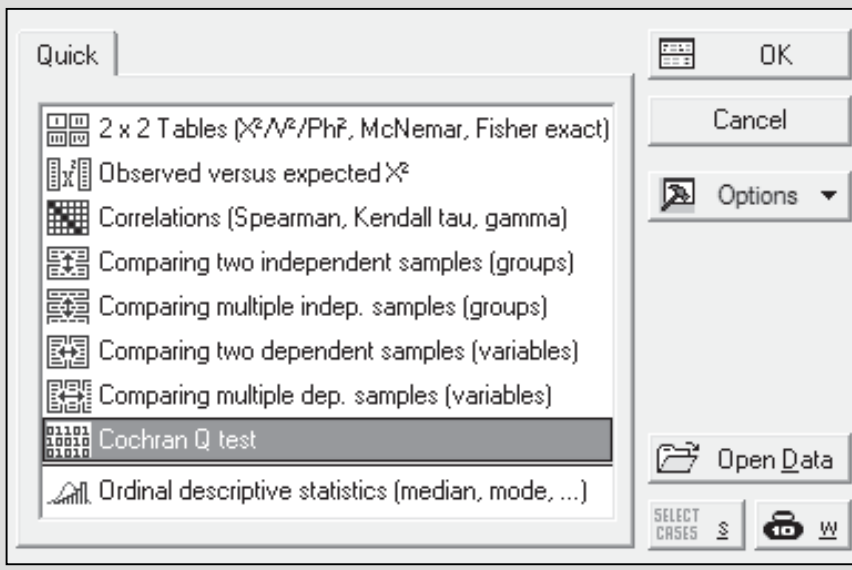
Paso 4. Se selecciona «Statistics» ⇨ «Nonparametrics».



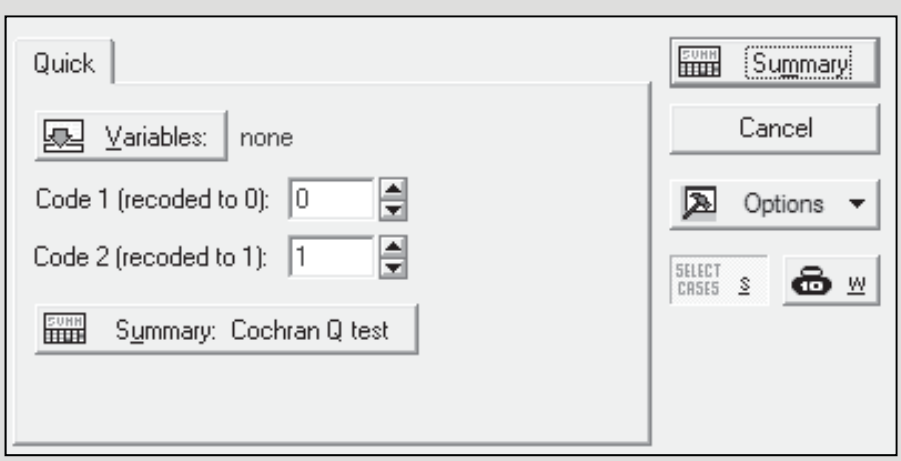
	1 Antibiótico				5 Día 20
22	0				0
23	0				0
24	0				0
25	0				0
26	0				0
27	0				1
28	0				0
29	0				0
30	0				0
31	0				0
32	0				1
33	0				0
34	0				0
35	0				0
36	0				1
37	0				0
38	0				0
39	0	0	0	0	0
40	0	0	0	0	0
41	0	0	0	0	0
42	0	0	0	0	0
43	0	0	0	0	0

CUADRO VII.5. *(Continuación)*[STATISTICA]

Paso 5. Seleccionamos el apartado «Cochran *Q* test».

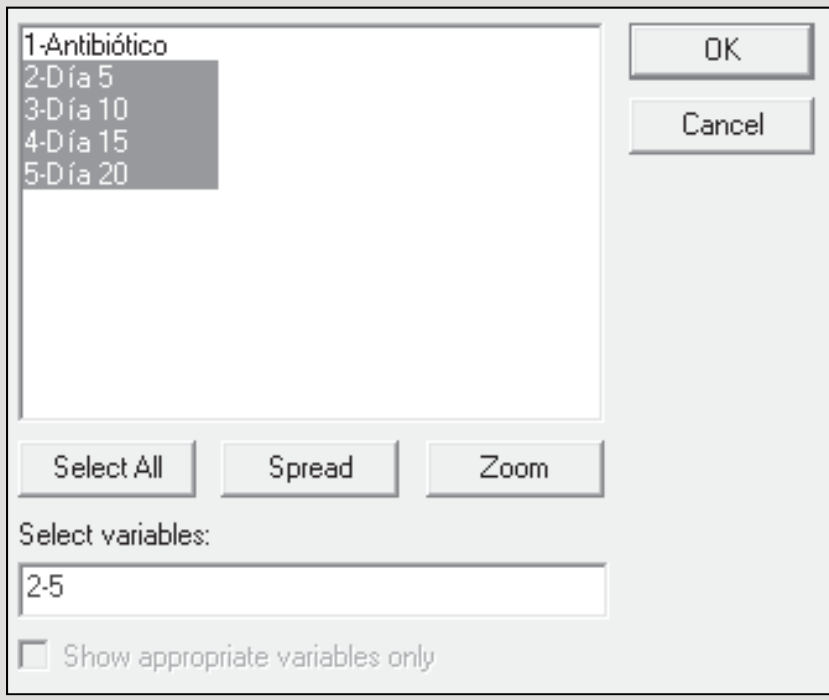


Paso 6. Esta prueba se calcula con valores 0 y 1. En caso de que nuestros datos no estén codificados con estos valores, STATISTICA se ocupa de recodificarlos.



Paso 7. En la ventana anterior debemos seleccionar las variables que queremos incluir en el análisis pulsando el icono «Variables». Aparece la siguiente ventana donde, en nuestro caso, se incluyen todos los días que se controló la respuesta alérgica de la piel al antibiótico.

CUADRO VII.5. *(Continuación)* STATISTICA



Paso 8. La tabla de resultados se muestra a continuación. Como podemos ver la probabilidad del estadístico *Q* con 3 grados de libertad es menor de 0,05 por lo que rechazamos la hipótesis nula de que las muestras son iguales, es decir, hay diferencias entre algunos de los días en los síntomas de alergia detectados en las personas sometidas al tratamiento con antibiótico (Test *Q* de Cochran = 50,4, *df* = 3, $p < 0{,}001$). Sin embargo, es importante resaltar que esto no significa que el antibiótico sea el responsable de la alergia, solo se demuestra que hay diferencias en el tiempo en la respuesta alérgica al antibiótico. Sería necesario también demostrar que la proporción de alergias es significativamente distinta del grupo no tratado con el antibiótico.

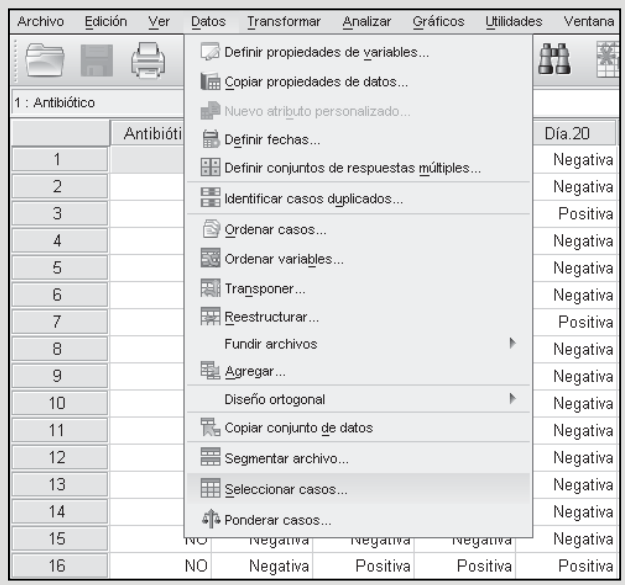
	Cochran Q Test (Cuadro VII.5.sta) Number of valid cases:50 Q = 50,40000, df = 3, p < ,000000 Include condition: v1=1		
Variable	Sum	Percent 0's	Percent 1's
Día 5	7,00000	86,00000	14,00000
Día 10	11,00000	78,00000	22,00000
Día 15	19,00000	62,00000	38,00000
Día 20	31,00000	38,00000	62,00000

CUADRO VII.5. *(Continuación)*[SPSS]

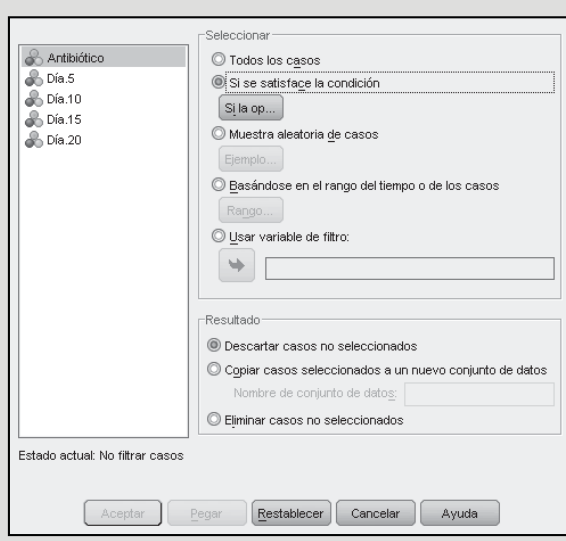
Aplicación con SPSS.

EJEMPLO. Vamos a utilizar los mismos datos que en el caso de la aplicación con R y con STATISTICA, los cuales se encuentran en el archivo **Cuadro VII.5.sav**. Sin embargo, en este caso vamos a analizar si existen diferencias significativas en la presencia de alergia entre los distintos días en el grupo de personas a las que no se les suministró el antibiótico.

Paso 1. Se selecciona «Datos» ⇨ «Seleccionar casos…».

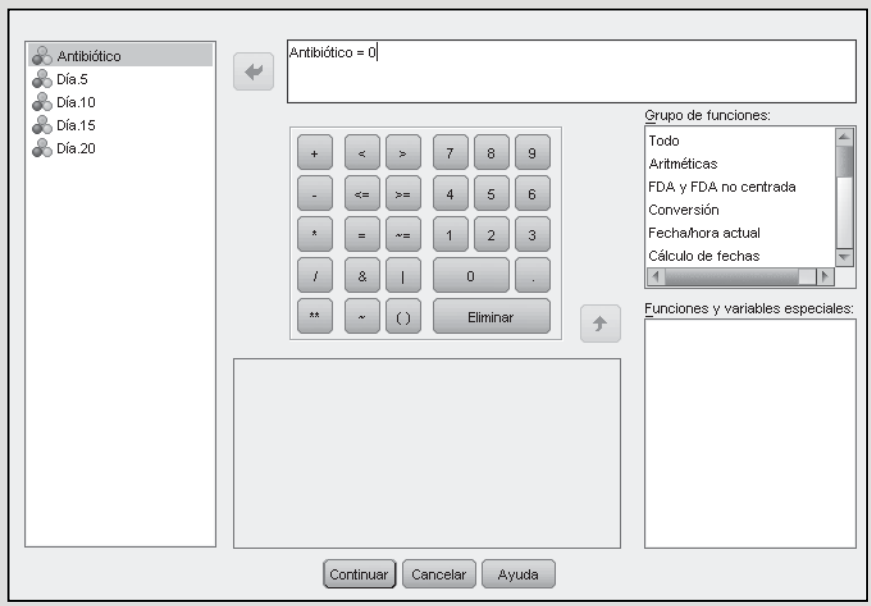


Paso 2. Se selecciona «Si se satisface la condición» y se pulsa en «Si la op…».

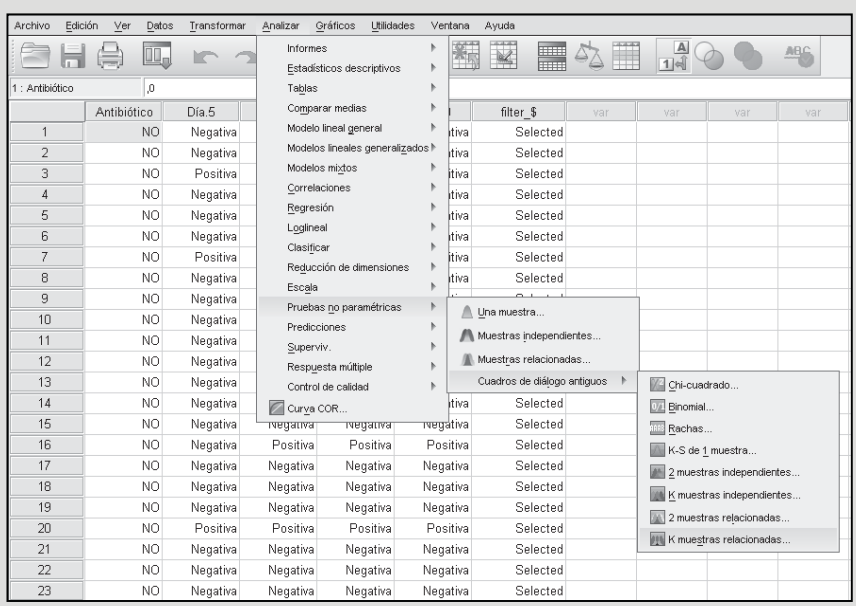


CUADRO VII.5. *(Continuación)*[SPSS]

Paso 3. Se seleciona la variable, en este caso «Antibiótico», y se pone la condición «Antibiótico=0», de tal forma que solo se trabaja con el grupo de personas que no recibieron el tratamiento. Después de pulsar «Continuar», cuando sale la ventana del paso 2 se pulsa «Aceptar».

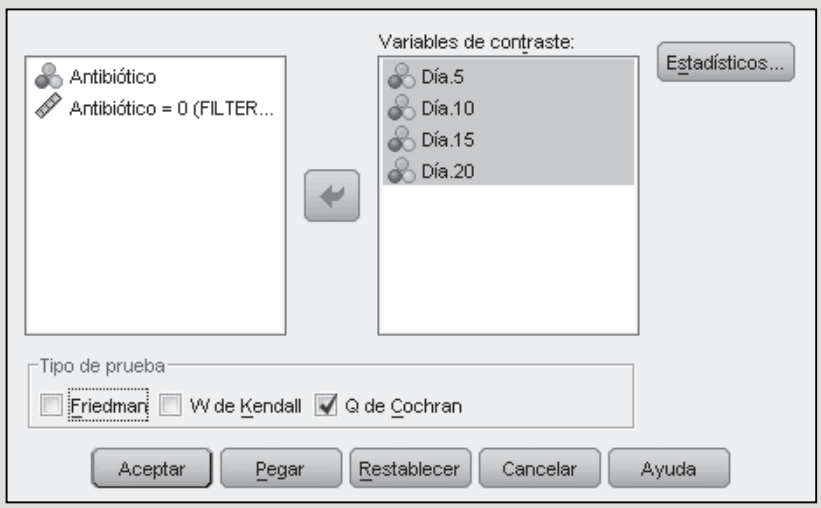


Paso 4. A continuación se selecciona «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «K muestras relacionadas…».



CUADRO VII.5. *(Continuación)*[SPSS]

Paso 5. En «Variables de contraste:» se incluyen las variables que se van a analizar, en este caso todos los días que se realizó el seguimiento, y en «Tipo de prueba» se selecciona «Q de Cochran».



Paso 6. La tabla de resultados se muestra a continuación. Como podemos ver la probabilidad del estadístico *Q* con 3 grados de libertad es menor que 0,05 por lo que rechazamos la hipótesis nula de que las muestras son iguales, es decir, también hay diferencias entre algunos de los días en los síntomas de alergia detectados en las personas que no fueron sometidas al tratamiento con antibiótico (Test *Q* de Cochran = 9,0. *df* = 3, $p < 0,029$). Sin embargo, como ya se comentó anteriormente en el ejemplo con STATISTICA para el caso del grupo al que se suministró el antibiótico, es necesario resaltar que esto no significa necesariamente que al cabo de los días el placebo generó una respuesta alérgica similar a la que se observó con el antibiótico, solo se demuestra que hay diferencias en el tiempo en la respuesta alérgica. Para poder comprobar esta posible relación entre tiempo y alergia es necesario hacer una prueba de independencia, lo cual se explica en el siguiente capítulo, en el que se analizarán los contrastes de independencia y asociación para variables cualitativas.

Estadísticos de contraste

N	50
Q de Cochran	9,000[a]
gl	3
Sig. asintót.	,029

a. 0 se trata como un éxito.

Capítulo VIII

CONTRASTES DE INDEPENDENCIA Y ASOCIACIÓN EN VARIABLES CUALITATIVAS

Uno de los principales propósitos de la estadística es determinar si existe asociación entre dos o más caracteres o variables de una misma muestra. Además de conocer si dos o más variables son independientes o dependientes es interesante conocer, también, cuál es la fuerza de dicha asociación.

De forma general diremos que la fórmula de aproximación para el estudio de la independencia es la medición de las diferencias entre las frecuencias observadas y las frecuencias que esperaríamos encontrar si las variables fuesen independientes.

VIII.1. Variables politómicas

VIII.1.1. PRUEBAS DE INDEPENDENCIA

VIII.1.1.1. Chi-cuadrado de Pearson y test *G* de razón de verosimilitud

Son los contrastes más utilizados, aunque no suelen emplearse con variables dicotómicas, ya que hay pruebas específicas para ellas.

Operativamente el cálculo de ambos estadísticos es igual al contraste de homogeneidad de muestras (véase apartados VII.1.1.1 y VII.1.1.2 del capítulo anterior), con la salvedad de que en este caso filas y columnas se refieren a variables diferentes de una misma muestra. Es decir, la tabla de contingencia sería la Tabla VII.1, que se mostró en el capítulo anterior, indicando las filas valores de una variable en lugar de muestras. Su cálculo se explica en el Cuadro VIII.1.

VIII.1.2. MEDIDAS DE ASOCIACIÓN

La medición de la intensidad de esta asociación puede estar basada en el cálculo de distancias entre variables o en la cantidad de variabilidad explicada si calculamos las frecuencias de una variable en función de la otra variable.

VIII.1.2.1. *V* de Cramer

El coeficiente *V* de Cramer es el siguiente:

$$V = \sqrt{\frac{\chi^2}{n(\min(f,c)-1)}}$$

siendo *f* y *c* el número de categorías en filas y columnas respectivamente y *n* el número de casos.

Un valor 0 indica independencia total de las dos variables, mientras que un valor 1 indica asociación perfecta. Tiene la ventaja de que permite analizar tablas de contingencia de cualquier tamaño. Su cálculo se explica en el Cuadro VIII.1.

VIII.1.2.2. Coeficiente de Contingencia

También es una medida asociada a χ^2. La fórmula es:

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}}$$

donde *n* es el número de elementos de la muestra.

Este coeficiente toma valores entre 0 y 1. Un valor próximo a cero indica independencia, y próximo a uno fuerte dependencia. Nunca puede tomar el valor límite 1, incluso cuando la dependencia es total. Su cálculo se explica en el Cuadro VIII.1.

VIII.1.2.3. Coeficiente de Incertidumbre

Es una medida de asociación basada en la reducción proporcional del error, con un rango entre 0 y 1, de las mediciones de una variable basándose en el conocimiento de la otra. Es decir, nos indica el porcentaje de la variabilidad explicado por la asociación de variables.

La tabla de contingencia que se utiliza es la Tabla VII.1, que se mostró en el capítulo anterior, aunque en lugar de muestras las filas se corresponden con una variable.

Existen dos métodos distintos para realizar esta prueba:

1. Considerando una variable independiente y otra dependiente. En este caso se pueden calcular dos estadísticos distintos ya que podemos considerar como variable dependiente tanto las filas como las columnas.

   Las medidas de asociación son, suponiendo la variable *Y* en columnas y *X* en filas:

$$I_{y|x} = \frac{(I_x + I_y - I_{xy})}{I_y} \quad \text{(columnas variable dependiente)}$$

$$I_{x|y} = \frac{(I_x + I_y - I_{xy})}{I_x} \quad \text{(filas variable dependiente)}$$

donde

$$I_Y = -\sum_{j=1}^{k} \frac{n_j}{n} \ln \frac{n_j}{n}$$

$$I_X = -\sum_{i=1}^{h} \frac{n_i}{n} \ln \frac{n_i}{n}$$

$$I_{XY} = -\sum_{i=1}^{h} \sum_{j=1}^{k} \frac{n_{ij}}{n} \ln \frac{n_{ij}}{n}$$

siendo h y k el número de filas y columnas, n el tamaño de la muestra, n_i y n_j las frecuencias marginales de fila y de columna, y n_{ij} las frecuencias conjuntas.

2. Sin considerar dependencia entre variables

$$I_{ind} = 2\frac{I_x + I_y - I_{xy}}{I_x + I_y}$$

Su cálculo se explica en el Cuadro VIII.1.

CUADRO VIII.1. Contraste de independencia con χ^2 y test *G* y las medidas de asociación *V* de Cramer y Coeficiente de Incertidumbre

Aplicación con R.

EJEMPLO. En el Cuadro VII.5 del capítulo anterior, que era un estudio para determinar si un antibiótico provoca una respuesta alérgica en la piel, el contraste de homogeneidad mostró que, tanto en el caso de las personas tratadas con el antibiótico como a las que no se les suministró, existían diferencias significativas en los síntomas de alergia detectados en la piel a lo largo del periodo de estudio. Ahora se quiere determinar si efectivamente existe una relación entre el tiempo y los síntomas de alergia. La variable respuesta al antibiótico es dicotómica (respuesta positiva o negativa), pero la variable tiempo no es dicotómica ya que hay datos de respuesta a los días 5, 10, 15 y 20 del estudio. Los datos están en el archivo **Cuadro VIII.1.V.csv**.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro VIII.1.R**, que se muestra en la siguiente ventana.

CUADRO VIII.1. *(Continuación)*[R]

Para que este *script* funcione correctamente es necesario instalar los paquetes "vcd" y "colorspace" como se explica en el Apéndice II.

```
# Cuadro VIII.1
# Contrastes de independencia, medidas de asociación y coeficiente de incertidumbre

#########################################################
# Sección modificable por el usuario
#########################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro VIII.1.V.csv",encoding="latin1")

# Selección de las variables de interés
# Si se quiere realizar la prueba de Friedman
# Las variables de interés deben seguir en estricto orden:
# Variable RESPUESTAS, FACTOR, BLOQUE
varInteres<-c("Respuesta","Día")

# Selección de las variables de agrupación.
#varAgrupa<-NULL
varAgrupa<-c("Antibiótico")

# Tipos de pruebas a realizar
#  1. Prueba chi-cuadrado.
#  2. Prueba G sin corrección de Williams.
#  3. Prueba G con corrección de Williams.
#  4. Prueba chi-cuadrado con corrección de Yates (sólo para tablas 2x2).
#  5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral.
#  6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor que.
#  7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor que.
#  8. Prueba ANOVA de Friedman.
#  9. Prueba de McNemar sin corrección de continuidad.
# 10. Prueba de McNemar con corrección de continuidad (sólo para tablas 2x2).
# 11. Prueba de Cochran.
# 12. Medidas de asociación.
# 13. Coeficiente de incertidumbre.
# 14. Pruebas de Cochran (sin corrección) (sólo tablas 2x2xk).
# 15. Pruebas de Cochran y Mantel-Haenszel (con corrección) (sólo tablas 2x2xk).
# 16. Cálculo de las razones de riesgo por estrato (sólo 2x2xk).
# 17. Pruebas Breslow & Day con la corrección de Tarone (sólo 2x2xk).

pruebas<-c(1,12,13)

# Nombre del archivo de salida con los resultados.
nomSalida<-"Salida Cuadro VIII.1.V.txt"
```

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro VIII.1.V.csv**.

datos<-read.csv2("Cuadro VIII.1.V.csv",encoding="latin1")

Paso 2. Se seleccionan las dos variables para las que se quiere determinar si existe asociación entre ellas, en este caso el grado de respuesta alérgica y el tiempo. No importa el orden en que se pongan las variables.

varInteres<-c("Respuesta","Día")

CUADRO VIII.1. *(Continuación)*^{R}

Paso 3. Es posible hacer los cálculos estadísticos agrupando las variables de la forma que se quiera. En el caso de seleccionar «NULL», no se realizaría ninguna agrupación y se calcularía considerando todos los datos de cada columna. En el ejemplo se quiere hacer la diferenciación entre las personas a las que se les suministró y las que no recibieron el antibiótico, por ello se pone «Antibiótico» como variable agrupación. Estas variables de agrupación nunca pueden coincidir con las variables de interés seleccionadas en el paso anterior.

```
#varAgrupa<-NULL
varAgrupa<-c("Antibiótico")
```

Paso 4. A continuación es necesario especificar los tipos de contrastes que se quieren realizar. En nuestro ejemplo, la prueba χ^2, las medidas de asociación y el coeficiente de incertidumbre.

```
#  1. Prueba chi-cuadrado.
#  2. Prueba G sin corrección de Williams.
#  3. Prueba G con corrección de Williams.
#  4. Prueba chi-cuadrado con corrección de Yates (sólo para tablas 2x2).
#  5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral.
#  6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor
que.
#  7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor
que.
#  8. Prueba ANOVA de Friedman.
#  9. Prueba de McNemar sin corrección de continuidad.
# 10. Prueba de McNemar con corrección de continuidad (sólo para tablas
2x2).
# 11. Prueba de Cochran.
# 12. Medidas de asociación.
# 13. Coeficiente de incertidumbre.
# 14. Pruebas de Cochran (sin corrección) (sólo tablas 2x2xk).
# 15. Pruebas de Cochran y Mantel-Haenszel (con corrección) (sólo tablas
2x2xk).
# 16. Cálculo de las razones de riesgo por estrato (sólo 2x2xk).
# 17. Pruebas Breslow & Day con la corrección de Tarone (sólo 2x2xk).

pruebas<-c(1,12,13)
```

Paso 5. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
nomSalida<-"Salida Cuadro VIII.1.V.txt"
```

Paso 6. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*.

CUADRO VIII.1. *(Continuación)*[R]

En el grupo al que no se suministró antibiótico, se acepta la hipótesis nula de independencia (χ^2_3 = 1,59, p = 0,661). Por tanto, a pesar de que el contraste de homogeneidad mostró diferencias significativas en la respuesta alérgica observada en la piel con el tiempo (véase Cuadro VII.5 en el ejemplo con R del capítulo anterior), estas diferencias observadas no son suficientes para establecer una dependencia significativa entre el tiempo y la respuesta alérgica en aquellos pacientes a los que se les suministró un placebo en vez del antibiótico.

Como no existe dependencia entre las variables, no tiene sentido analizar los coeficientes de asociación.

```
$NO
$NO$tabla
          Día
Respuesta  Día.10 Día.15 Día.20 Día.5
  Negativa     43     43     42    46
  Positiva      7      7      8     4

$NO$listaPruebas
$NO$listaPruebas[[1]]

        Pearson's Chi-squared test

data:  tbl1
X-squared = 1.5915, df = 3, p-value = 0.6613


$NO$listaPruebas[[2]]
                    X^2 df P(> X^2)
Likelihood Ratio 1.7181  3  0.63293
Pearson          1.5915  3  0.66132

Phi-Coefficient   : 0.089
Contingency Coeff.: 0.089
Cramer's V        : 0.089

$NO$listaPruebas[[3]]
Uncertainty coefficents or Theil's Uncertainty
X:   0.003098296
Y:   0.01111620
X|Y: 0.004845938
```

Por el contrario, en el grupo de personas que recibieron antibiótico, se rechaza la hipótesis nula de independencia (χ^2_3 = 29,9, p < 0,001), es decir, existe una relación significativa entre el tiempo y la respuesta alérgica observada en la piel.

CUADRO VIII.1. *(Continuación)*[R]

Entre los coeficientes de asociación consideramos el *V* de Cramer, el coeficiente de incertidumbre, y el coeficiente de contingencia. No consideramos phi ya que la tabla no es 2x2.

El coeficiente *V* de Cramer es 0,39 y el de contingencia 0,36, lo que indica asociación débil. El coeficiente de incertidumbre *y*-dependiente es 0,1186, es decir, la variable *x* (el tiempo) reduce un 11,86% la incertidumbre sobre la predicción de la variable *y* (la alergia). Naturalmente consideramos que la variable alergia es dependiente del tiempo, y no al revés.

Con estos resultados podemos concluir que nuestras variables están relacionadas significativamente con valor de 0,39 y, por tanto, como la alergia (variable *y*) depende del tiempo (variable *x*), se puede concluir que el tiempo explica un 12% de la alergia observada en la piel en las personas a las que se les suministró el antibiótico.

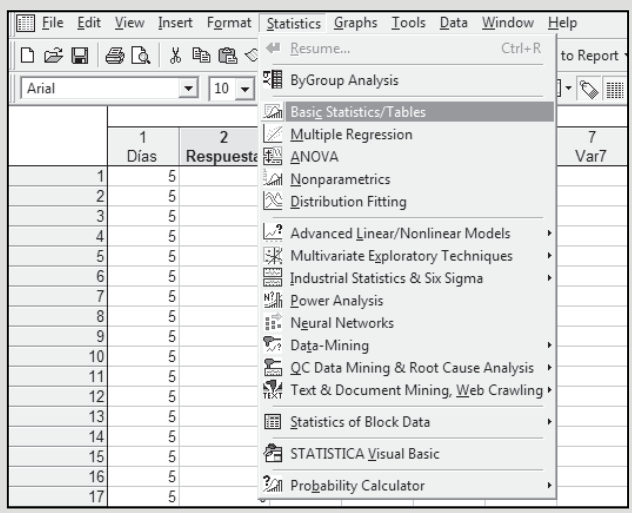
```
$SI
$SI$tabla
          Día
Respuesta  Día.10 Día.15 Día.20 Día.5
  Negativa     39     31     19    43
  Positiva     11     19     31     7

$SI$listaPruebas
$SI$listaPruebas[[1]]

        Pearson's Chi-squared test

data:  tbl1
X-squared = 29.9465, df = 3, p-value = 1.416e-06


$SI$listaPruebas[[2]]
                    X^2 df   P(> X^2)
Likelihood Ratio 30.414  3 1.1291e-06
Pearson          29.947  3 1.4163e-06

Phi-Coefficient   : 0.387
Contingency Coeff.: 0.361
Cramer's V        : 0.387

$SI$listaPruebas[[3]]
Uncertainty coefficents or Theil's Uncertainty
X:   0.05484806
Y:   0.1186136
X|Y: 0.07501054
```

CUADRO VIII.1. *(Continuación)*[STATISTICA]

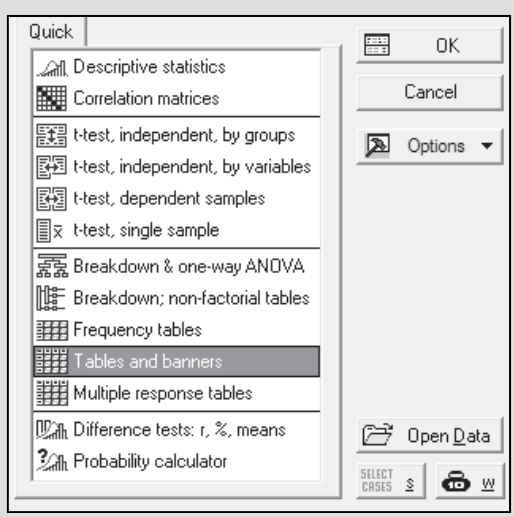
Aplicación con STATISTICA.

EJEMPLO. Se va a utilizar el mismo ejemplo que se usó con R, es decir, determinar si existe relación entre la respuesta alérgica a un antibiótico y el tiempo que trascurre después de que se suministró el antibiótico. Como en el ejemplo del contraste de homogeneidad con STATISTICA se utilizó el grupo de personas a las que se les había suministrado el antibiótico (Cuadro VII.5), en este ejemplo haremos lo mismo. Los datos están en el archivo **Cuadro VIII.1.sta**.

Paso 1. Se selecciona «Statistics» ⇨ «Basic Statistics/Tables».

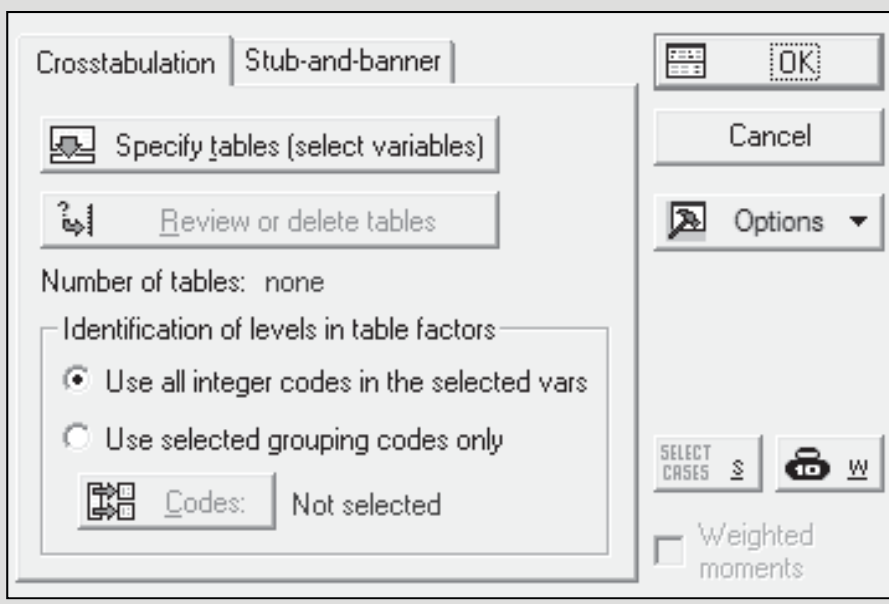


Paso 2. A continuación se selecciona «Tables and banners».

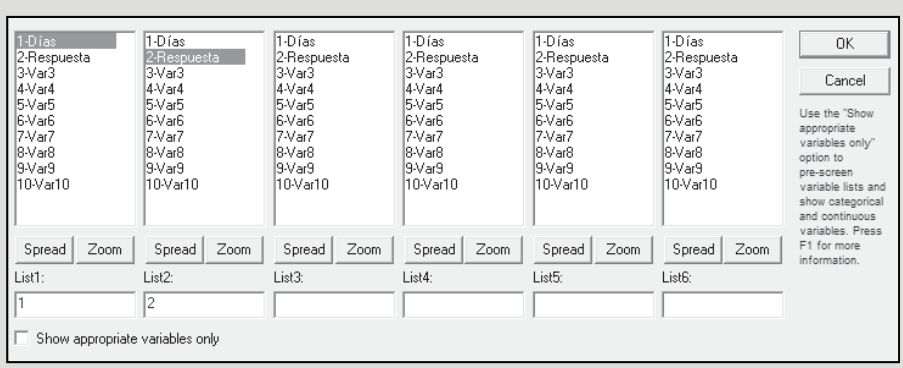


CUADRO VIII.1. *(Continuación)* STATISTICA

Paso 3. Activamos la pestaña «Crosstabulation» (tabla de contingencia) donde se nos permite seleccionar las categorías con las que queremos trabajar («Use selected grouping codes only») o usar todas las categorías de los datos («Use all integer codes in the selected vars»), que es la opción que trae por defecto el programa. Nosotros trabajaremos con todas las categorías y no modificaremos este punto. Pulsamos «Specify tables (select variables)» para seleccionar las variables que se van a estudiar.

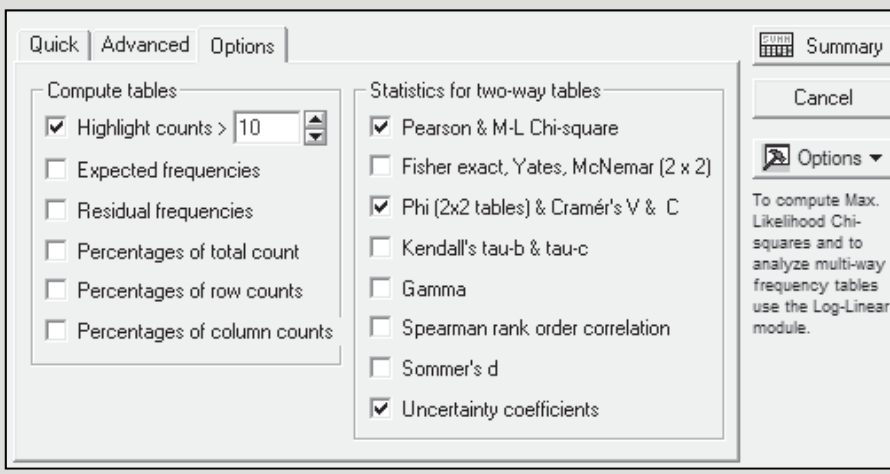


Paso 4. Se seleccionan las variables «Días» y «Respuesta». No importa el orden en que se pongan las variables.

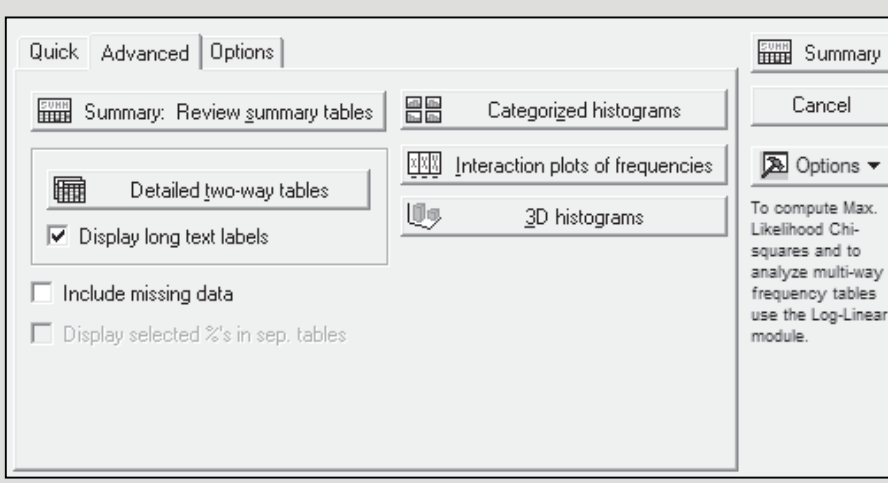


CUADRO VIII.1. *(Continuación)*[STATISTICA]

Paso 5. Pulsamos «OK» en las dos pantallas hasta que aparece la siguiente ventana, donde hay tres pestañas diferentes, de las cuales dos son de resultados, «Quick» y «Advanced», y una para seleccionar los estadísticos «Options». En la pestaña «Options» seleccionamos la opción «Pearson & ML Chi-square». Estos 2 estadísticos se corresponden con el test χ^2 (Pearson Chi-square) y el test *G* (Maximum Likelihood Chi-square), respectivamente. También se selecciona «Phi (2x2 tables) & Cramer´s V & C» y «Uncertainty coefficients» para calcular los coeficientes de asociación.



Paso 6. En la pestaña "Advanced" aparece la siguiente ventana, donde pulsando el icono "Summary: Review summary tables" aparece la tabla de contingencia de nuestros resultados.



CUADRO VIII.1. *(Continuación)*[STATISTICA]

Summary Frequency Table (Cuadro VIII.1.sta) Marked cells have counts > 10 (Marginal summaries are not marked)			
Días	Respuesta 0	Respuesta 1	Row Totals
5	43	7	50
10	39	11	50
15	31	19	50
20	19	31	50
All Grps	132	68	200

Paso 7. En la ventana del paso 6 pulsando el icono "Detailed two-way tables" aparecen los resultados.

	Statistics: Días(4) x Respuesta(2) (Cuadro VIII.1.sta)		
Statistic	Chi-square	df	p
Pearson Chi-square	29,94652	df=3	p=,00000
M-L Chi-square	30,41422	df=3	p=,00000
Phi	,3869530		
Contingency coefficient	,3608775		
Cramér's V	,3869530		
Uncertainty coefficient	X=,0548481	Y=,1186136	X|Y=,07501

En la tabla aparece una primera fila que contiene el estadístico χ^2 y una segunda con el test *G.* En ambos casos, se nos indica el valor del estadístico, los grados de libertad y la probabilidad del contraste. La probabilidad es menor de 0,001 para los 2 contrastes. Por lo tanto, se rechaza la hipótesis nula de independencia, es decir, existe una relación significativa entre el tiempo y la respuesta alérgica observada en la piel (χ^2_3 = 29,9, p < 0,001 y el test G = 30,4, df = 3, p < 0,001).

Entre los coeficientes de asociación consideramos el *V* de Cramer, el coeficiente de incertidumbre, y el coeficiente de contingencia. No consideramos phi ya que la tabla no es 2x2.

El coeficiente *V* de Cramer es 0,39 y el de contingencia 0,36, lo que indica asociación débil. El coeficiente de incertidumbre *y*-dependiente es 0,1186, es decir, la variable *x* (el tiempo) reduce un 11,86% la incertidumbre sobre la predicción de la variable *y* (la alergia). Naturalmente consideramos que la variable alergia depende del tiempo, y no al revés, por lo que no tiene sentido interpretar el coeficiente x-dependiente.

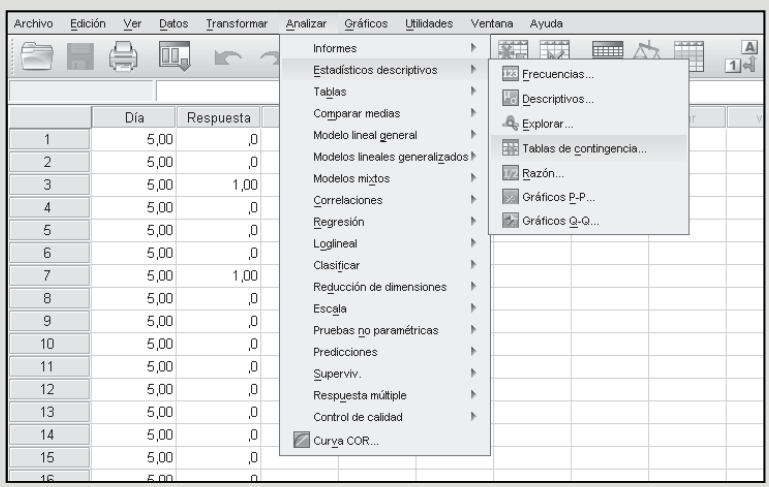
Con estos resultados podemos concluir que nuestras variables están relacionadas significativamente, aunque la asociación es débil: el paso del tiempo influye parcialmente en la alergia observada en los pacientes a los que se suministró antibiótico.

CUADRO VIII.1. *(Continuación)*[SPSS]

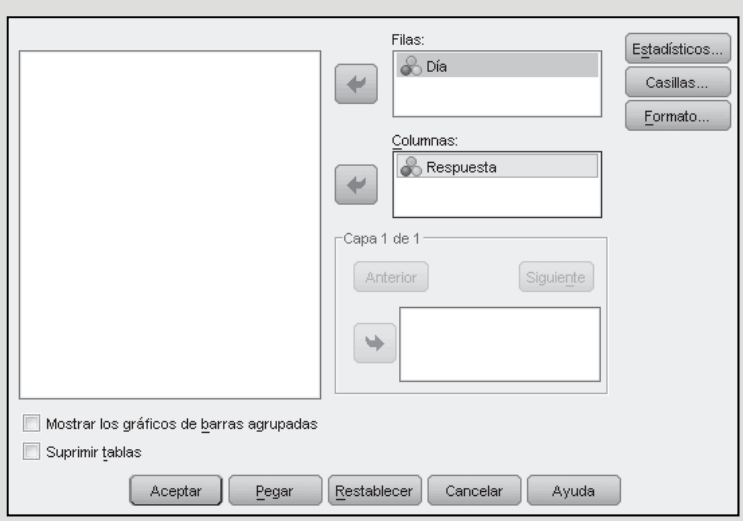
Aplicación con SPSS.

EJEMPLO. Utilizaremos los mismos datos que con R y STATISTICA, analizando el grupo de pacientes a los que no se había suministrado el antibiótico (Cuadro VII.5 del capítulo anterior). Los datos están en el archivo **Cuadro VIII.1.sav**.

Paso 1. Los datos hay que organizarlos como se ve en la siguiente pantalla y en el archivo **Cuadro VIII.1.sav**. Se selecciona «Analizar» ⇨ «Estadísticos descriptivos» ⇨ «Tablas de contingencia…».

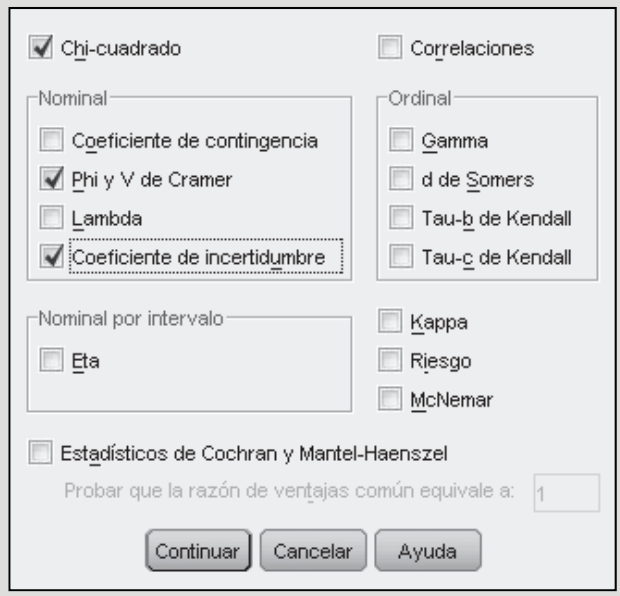


Paso 2. Las variables a comparar se ponen en «Filas» y «Columnas» sin importar el orden.



CUADRO VIII.1. *(Continuación)*[SPSS]

Paso 3. En la ventana anterior hay que pulsar en «Estadísticos…» y en la ventana que aparece marcar «Chi-cuadrado», «Phi y V de Cramer» y «Coeficiente de incertidumbre».



Paso 4. Los resultados de la pruebas de independencia se muestran en la siguiente tabla. Tanto en el test χ^2 (χ^2_3 = 1,59, p = 0,661) como en el test G (G = 1,72, df = 3, p = 0,633) se acepta la hipótesis nula de que existe independencia. Por tanto, a pesar de que el contraste de homogeneidad mostró diferencias significativas en la respuesta alérgica observada en la piel con el tiempo (véase Cuadro VII.5 en el ejemplo con SPSS), estas diferencias observadas no son suficientes para establecer una dependencia significativa entre el tiempo y la respuesta alérgica en aquellos pacientes a los que se les suministró un placebo en vez del antibiótico.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	1,592[a]	3	,661
Razón de verosimilitudes	1,718	3	,633
Asociación lineal por lineal	1,267	1	,260
N de casos válidos	200		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 6,50.

Como no existe dependencia entre las variables, no tiene sentido analizar los coeficientes de asociación. Para ver cómo se interpretan estos coeficientes en una situación en la que sí exista una dependencia significativa, véase el caso explicado para STATISTICA.

VIII.2. Variables dicotómicas

VIII.2.1. PRUEBAS DE INDEPENDENCIA

VIII.2.1.1. Corrección de Yates y prueba de Fisher

Para testar la independencia en variables dicotómicas, cuando solo se tiene una tabla 2x2, se puede usar el test χ^2 con la corrección de Yates, el cual ya se explicó en el apartado VII.1.2.1 del capítulo anterior.

La prueba de Fisher, la cual se explicó brevemente en el apartado VII.1.2.2 del capítulo anterior, también se utiliza cuando la tabla de contingencia es 2x2 y debe usarse en vez de la prueba χ^2 cuando en la tabla de contingencia exista alguna celda con frecuencias esperadas menores de 5.

La corrección de Yates y la prueba de Fisher se explican para el caso de una prueba de independencia en el Cuadro VIII.2.

VIII.2.1.2. Pruebas de Cochran y Mantel-Haenszel

Al igual que el test de Fisher, las pruebas de Cochran (1954) y Mantel & Haenszel (1959) también sirven para testar la independencia en una tabla de contingencia 2x2. Sin embargo, tienen la ventaja sobre la prueba de Fisher de que también se pueden usar cuando existan varias tablas 2x2. Estas tablas de contingencia son del tipo *K*x2x2, es decir, hay tantas tablas 2x2 como *K* estratos. En estas tablas, se trata de estudiar si existe o no asociación entre dos variables dicotómicas, cuando se dispone de información referida a varios estratos. Por ejemplo, se quiere determinar si la presencia de cáncer de cuello uterino está relacionada con la presencia del virus del papiloma humano en mujeres de diferentes edades o por el contrario son independientes. Por lo tanto, tendríamos varias tablas 2x2 (presencia o no presencia de cáncer y presencia o no presencia del virus, ambas variables dicotómicas) en varios estratos que serían las diferentes clases de edad en que se divida el grupo de mujeres estudiadas. En este ejemplo, analizar cada tabla 2x2 por separado con la prueba de Fisher no daría información sobre el efecto de la edad.

El estadístico de Cochran es el siguiente:

$$\chi^2_{Cochran} = \frac{\left(\sum_k n_k - \sum_k m_k\right)^2}{\sum_k \sigma^2_{n_k}}$$

donde:

k = cada uno de los estratos

n_k = frecuencia observada en cualquiera de las casillas del estrato k (solo una y siempre es la misma en todos los estratos)

m_k = frecuencia esperada correspondiente a n_k

$$\sigma^2_{n_k} = \frac{n_{f1k} n_{f2k} n_{c1k} n_{c2k}}{n^3}$$

n_{f1k}, n_{f2k}, n_{c1k} y n_{c2k} son las cuatro frecuencias marginales asociadas a las tablas 2x2 de cada estrato *k*.

El estadístico Mantel-Haenszel es muy parecido al de Cochran, excepto por dos aspectos:

1) Utiliza la corrección por continuidad (resta medio punto al numerador antes de elevar al cuadrado).
2) Cambia ligeramente el denominador de la varianza, utilizando $n^2(n-1)$ en lugar de n^3.

En ambos estadísticos, el de Cochran y Mantel-Haenszel, si el valor de *p* es mayor que 0,05, se acepta la hipótesis de independencia y se concluye que, una vez tenido en cuenta el efecto de los estratos, las dos variables dicotómicas no están asociadas.

El cálculo de estas pruebas se explica en el Cuadro VIII.2.

VIII.2.2. MEDIDAS DE ASOCIACIÓN

En el caso de tablas de contingencia 2x2 se puede usar cualquiera de los tres coeficientes explicados anteriormente en el apartado VIII.1.2: *V* de Cramer, Coeficiente de Contingencia y Coeficiente de Incertidumbre. Además de estos tres coeficientes se puede usar el coeficiente Phi.

VIII.2.2.1. Phi (ϕ)

El rango de este coeficiente solo es entre 0 y 1 en las tablas de contingencia 2x2 y, por ello, solo se debe usar cuando las variables a comparar son ambas dicotómicas.

Phi (ϕ) es una medida basada en el estadístico χ^2. La fórmula es:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

donde *n* es el número de elementos de la muestra.

El cálculo de este coeficiente se explica en el Cuadro VIII.2.

VIII.2.2.2. Riesgo relativo, razón de ventajas (*odds ratio*) y las pruebas de Breslow-Day y de Tarone

Cuando tenemos tablas de contingencia *K*x2x2, de las que se habló en el apartado anterior, la importancia de una asociación se estima mejor usando el riesgo relativo y la razón de ventajas (*odds ratio*). Estas medidas se pueden usar también en tablas 2x2, al igual que los coeficientes descritos anteriormente (*V* de Cramer, Coeficiente de Contingencia, Coeficiente de Incertidumbre y Phi), pero en el caso de las tablas *K*x2x2 estas medidas tienen la ventaja de que se pueden comparar entre los estratos.

Riesgo relativo nos indica cuánto más probable es que ocurra el suceso de un primer grupo frente al segundo. En una tabla 2x2 como la que se muestra en la

Tabla VIII.1 el riesgo o probabilidad del suceso considerado en el grupo A y en el grupo B serían:

Tabla VIII.1. Tabla de contingencia 2x2.

Variable 1	Variable 2		
	Grupo A	***Grupo B***	
			Total
Suceso	*a*	*c*	a+c
No suceso	*b*	*d*	b+d
Total	a+b	c+d	*n*

$$Riesgo(A) = \frac{a}{(a+b)} \qquad Riesgo(B) = \frac{c}{(c+d)}$$

el cociente de ambas cantidades es el Riesgo relativo del suceso (RR_{suceso}) y del mismo modo se podría calcular el Riesgo relativo del suceso contrario o "no suceso" ($RR_{no\ suceso}$):

$$RR_{suceso} = \frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}} \qquad RR_{no\_suceso} = \frac{\frac{b}{(a+b)}}{\frac{d}{(c+d)}}$$

El cociente de ambos riesgos relativos es igual a la Razón de ventajas o Razón de predominios (*OR*, Odds Ratio):

$$OR = \frac{RR_{suceso}}{RR_{no\_suceso}} = \frac{(a*d)}{(b*c)}$$

La interpretación del Riesgo relativo es más intuitiva que la de la Razón de predominios. A pesar de ello la razón de ventajas se usa mucho, principalmente por su relación con la Regresión Logística, que se verá en otro capítulo, con importantes aplicaciones en Epidemiología y otras áreas.

La prueba de Breslow & Day (1980) y la modificación de esta prueba de Tarone (1985) son pruebas de homogeneidad que permiten comparar las razones de ventaja entre los distintos estratos, en tablas de contingencia *K*x2x2. Con esta información se consigue determinar si el grado de asociación entre las dos variables dicotómicas varía en función de los diferentes estratos. En el ejemplo que se mencionó anteriormente, nos informaría acerca de si la posible asociación entre el virus del papiloma humano y el cáncer de cuello uterino aumenta o se hace menos intensa con la edad de las mujeres.

El cálculo del riesgo relativo, la razón de ventajas y las pruebas de Breslow-Day y Tarone se explican en el Cuadro VIII.2.

CUADRO VIII.2. Contrastes de independencia y medidas de asociación para tablas de contingencia de 2x2

Aplicación con R.

EJEMPLO. En el archivo **Cuadro VIII.2.V.csv** hay información de una muestra de personas sobre el salario bruto mensual (1 es menos de 1.000 €, 2 es entre 1.000 y 1.499 €, 3 es entre 1.500 y 2.000 € y 4 es más de 2.000 €), si tienen o no tienen ya contratada una línea de internet con una velocidad de menos de 5 Megas (0 ya tienen y 1 no tienen internet) y de estas personas cuantas contratan la línea de internet -que ofrece la empresa- de más de 10 Megas (0 contratan la nueva línea y 1 no quieren la oferta que ofrece la empresa). El objetivo de la empresa es poder aumentar el número de contrataciones y, por ello, se plantea si el que acepten o no acepten la oferta depende de que tengan o no tengan internet y/o del salario bruto mensual.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro VIII.2.R**, que se muestra en la siguiente ventana. Para que este *script* funcione correctamente es necesario instalar el paquete "vcd" como se explica en el Apéndice II.

```
# Cuadro VIII.2
# Pruebas de Cochran y Mantel-Haenszel

########################################################
# Sección modificable por el usuario
########################################################
# Lectura de la base de datos
datos<-read.csv2("Cuadro VIII.2.V.csv",encoding="latin1")

# Selección de las variables de interés
# Si se quiere realizar la prueba de Friedman
# Las variables de interés deben seguir en estricto orden:
# Variable respuestas, Factor, bloque
varInteres<-c("Internet","Contratación" )

# Selección de las variables de agrupación.
varAgrupa<-NULL

# Para el caso de las pruebas de Cochran y Mantel-Haenszel (14 y 15)
# Se coloca la variable de estrato.
varEstratos<-c("Nivel.económico")

# Tipos de pruebas a realizar
#  1. Prueba chi-cuadrado.
#  2. Prueba G sin corrección de Williams
#  3. Prueba G con corrección de Williams
#  4. Prueba chi-cuadrado con corrección de Yates (Sólo para tablas 2x2).
#  5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral
#  6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor que.
#  7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor que.
#  8. Prueba ANOVA de Friedman
#  9. Prueba de McNemar sin corrección de continuidad.
# 10. Prueba de McNemar con corrección de continuidad (Sólo para tablas 2x2).
# 11. Prueba de Cochran.
# 12. Medidas de asociación.
# 13. Coeficiente de incertidumbre.
# 14. Pruebas de Cochran (sin corrección) (solo tablas 2x2xk)
# 15. Pruebas de Cochran y Mantel-Haenszel (con corrección) (sólo tablas 2x2xk)
# 16. Cálculo de las razones de riesgo por estrato (solo 2x2xk)
# 17. Pruebas Breslow & Day con la corrección de Tarone. (solo 2x2xk)

pruebas<-c(1,4,14,15,16,17)

# Nombre del archivo de salida con los resultados.
nomSalida<-"Salida Cuadro VIII.2.V.txt"
```

CUADRO VIII.2. *(Continuación)*[R]

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro VIII.2.V.csv**.

```
datos<-read.csv2("Cuadro VIII.2.V.csv",encoding="latin1")
```

Paso 2. Se seleccionan las dos variables cuya asociación se quiere determinar. No importa el orden en que se pongan las variables.

```
varInteres<-c("Internet","Contratación" )
```

Paso 3. En este caso, no existe variable de agrupación.

```
varAgrupa<-NULL
```

Paso 4. Para realizar estos tests es necesario definir estratos en las variables, en este caso, el nivel económico.

```
varEstratos<-c("Nivel.económico")
```

Paso 5. A continuación es necesario especificar los tipos de contrastes que se quieren realizar.

```
#  1. Prueba chi-cuadrado.
#  2. Prueba G sin corrección de Williams.
#  3. Prueba G con corrección de Williams.
#  4. Prueba chi-cuadrado con corrección de Yates (sólo para tablas 2x2).
#  5. Prueba "exacta" de Fisher (Sólo para tablas 2x2), bilateral.
#  6. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: menor
que.
#  7. Prueba "exacta" de Fisher (Sólo para tablas 2x2), unilateral: mayor
que.
#  8. Prueba ANOVA de Friedman.
#  9. Prueba de McNemar sin corrección de continuidad.
# 10. Prueba de McNemar con corrección de continuidad (sólo para tablas
2x2).
# 11. Prueba de Cochran.
# 12. Medidas de asociación.
# 13. Coeficiente de incertidumbre.
# 14. Pruebas de Cochran (sin corrección) (sólo tablas 2x2xk).
# 15. Pruebas de Cochran y Mantel-Haenszel (con corrección) (sólo tablas
2x2xk).
# 16. Cálculo de las razones de riesgo por estrato (sólo 2x2xk).
# 17. Pruebas Breslow & Day con la corrección de Tarone (sólo 2x2xk).

pruebas<-c(1,4,14,15,16,17)
```

Paso 6. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
nomSalida<-"Salida Cuadro VIII.2.V.txt"
```

CUADRO VIII.2. *(Continuación)*[R]

Paso 7. Se selecciona todo el archivo con *ctrl-A* y luego *ctrl-R* para ejecutar el *script*.

Paso 8. Ahora podemos abrir el archivo de salida y, en primer lugar, tenemos una serie de tablas de doble entrada con los datos resumidos.

```
$tabla
, , Nivel.económico = 1000 a 1499 €

        Contratación
Internet   No   Si
      No  156  720
      Si  146 1366

, , Nivel.económico = 1500 a 2000 €

        Contratación
Internet   No   Si
      No   66  420
      Si   36  598

, , Nivel.económico = más de 2000 €

        Contratación
Internet   No   Si
      No  113  872
      Si   45  688

, , Nivel.económico = Menos de 1000 €

        Contratación
Internet   No   Si
      No  110  311
      Si   90  663
```

A continuación podemos ver una serie de resultados de los diferentes análisis realizados. El χ^2 sin (χ^2_1 = 80,87, p < 0,001) y con corrección de Yates (χ^2_1 = 80,17, p < 0,001) son significativos, por lo que concluimos que existe asociación entre ambas variables, lo que quiere decir que existe relación entre tener o no tener internet y contratar la nueva línea de más alta velocidad.

```
$listaPruebas
$listaPruebas[[1]]

        Pearson's Chi-squared test

data:  tbl1
X-squared = 80.8766, df = 1, p-value < 2.2e-16


$listaPruebas[[2]]

        Pearson's Chi-squared test with Yates' continuity correction

data:  tbl1
X-squared = 80.1775, df = 1, p-value < 2.2e-16
```

CUADRO VIII.2. *(Continuación)*[R]

Las pruebas de Cochran (sin corrección de continuidad, $\chi^2_1 = 103,5$, $p < 0,001$) y Mantel-Haenszel (con corrección de continuidad, $\chi^2_1 = 102,7$, $p < 0,001$) también son significativas, con lo cual también nos indican la existencia de una relación de dependencia entre las variables.

También se nos muestran los valores de razón de ventajas (*odds ratio*), que están relacionados, en este caso, con la probabilidad de contratación cuando ya se tiene una línea de internet. Para una persona que gana más de 2.000 € al mes (nivel 4 en nuestro ejemplo), la razón de ventajas, que cuantifica el efecto sobre el alta en la empresa de tener ya una línea de internet, es 2,606. Para una persona que gana menos de 1.000 € (nivel 1) es 2,027.

```
$listaPruebas[[3]]

        Mantel-Haenszel chi-squared test without continuity correction

data:  tbl1
Mantel-Haenszel X-squared = 103.514, df = 1, p-value < 2.2e-16
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.901584 2.603403
sample estimates:
common odds ratio
         2.224992


$listaPruebas[[4]]

        Mantel-Haenszel chi-squared test with continuity correction

data:  tbl1
Mantel-Haenszel X-squared = 102.7047, df = 1, p-value < 2.2e-16
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.901584 2.603403
sample estimates:
common odds ratio
         2.224992


$listaPruebas[[5]]

        Cochran-Mantel-Haenszel Chi-square Test

data:  tbl1
CMH statistic = 103.514, df = 1.000, p-value = 0.000, MH Estimate =
2.225, Pooled Odd Ratio = 2.003, Odd Ratio of level 1 = 2.027, Odd
Ratio of level 2 = 2.610, Odd Ratio of level 3 = 1.981, Odd Ratio of
level 4 = 2.606
```

El test de Breslow y la correción de Tarone sirven para comprobar si hay diferencias significativas entre estratos. En nuestro caso, el valor de *p* asociado a estos estadísticos es 0.474, con lo cual aceptamos la hipótesis nula de que no existen diferencias entre estratos. Esto quiere decir que el nivel económico no influye en la decisión de contratar la línea de internet.

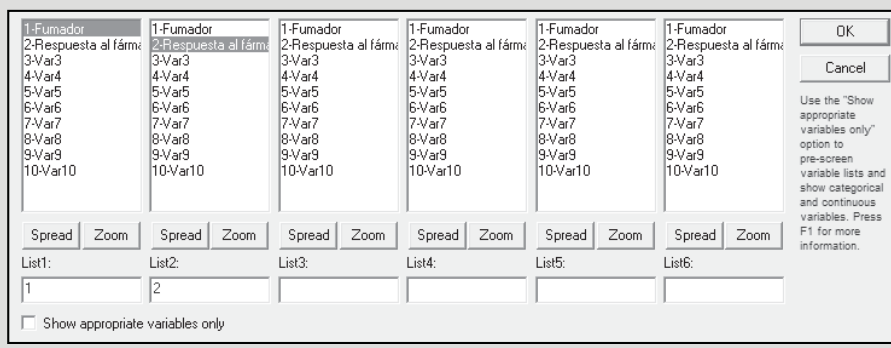
```
$listaPruebas[[6]]
Breslow and Day test (with Tarone correction):
Breslow-Day X-squared          = 2.509395
Breslow-Day-Tarone X-squared   = 2.50513

Test for test of a common OR: p-value =  0.4743647
```

CUADRO VIII.2. *(Continuación)*[STATISTICA]

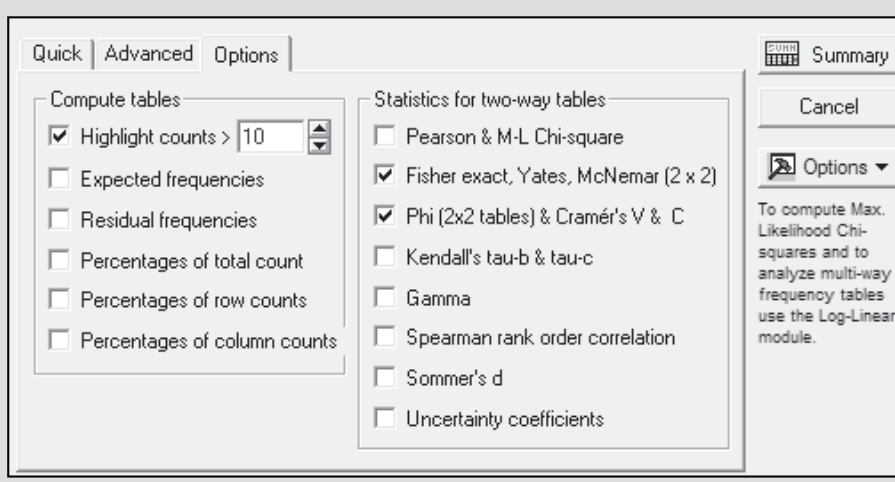
Aplicación con STATISTICA.

EJEMPLO. El objetivo es determinar si existe relación entre el tabaco (0 no fuma y 1 si fuma) y la respuesta que tiene un grupo de personas a un fármaco encargado de reducir la hipertensión (0 el fármaco no reduce la hipertensión y 1 el fármaco si reduce la hipertensión). Los datos están en el archivo **Cuadro VIII.2.sta**.

Paso 1. Se siguen los pasos 1 a 3 del Cuadro VIII.1 y aparece la ventana que se muestra a continuación, donde se introducen las dos variables a estudiar.

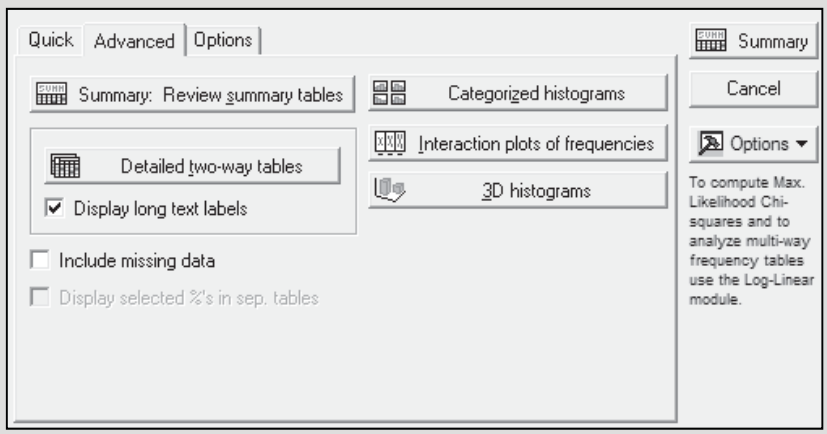


Paso 2. Después de pulsar dos veces «OK» aparece la ventana que está a continuación, donde hay tres pestañas diferentes, de las cuales dos son de resultados, «Quick» y «Advanced», y una para seleccionar los estadísticos «Options». En la pestaña «Options» seleccionamos la opción «Fisher exact, Yates, McNemar (2 x 2)» y «Phi(2x2 tables) & Cramer´s V & C».



CUADRO VIII.2. *(Continuación)*[STATISTICA]

Paso 3. En la ventana de la pestaña «Advanced» seleccionamos el icono «Detailed two-ways table» para que el programa nos muestre el resultado de los estadísticos solicitados.



	Statistics: Fumador(2) x Respuesta al fármaco(2) (Cuadro VIII.2.sta)		
Statistic	Chi-square	df	p
Pearson Chi-square	168,1429	df=1	p=0,0000
M-L Chi-square	190,6185	df=1	p=0,0000
Yates Chi-square	164,8610	df=1	p=0,0000
Fisher exact, one-tailed			p=0,0000
two-tailed			p=,00000
McNemar Chi-square (A/D)	11,60744	df=1	p=,00066
(B/C)	6,035714	df=1	p=,01402
Phi for 2 x 2 tables	,7891461		
Tetrachoric correlation	,9529828		
Contingency coefficient	,6194859		

Como en la tabla de contingencia no hay más de una celda con la frecuencia menor que 5, se puede usar la prueba de Yates. En la prueba χ^2 con la corrección de Yates la $p < 0{,}001$ y, por tanto, se rechaza la hipótesis nula de independencia, es decir, existe una relación significativa entre que la persona sea fumador o no fumador y la respuesta al fármaco para reducir hipertensión ($\chi^2_1 = 164{,}9$, $p < 0{,}001$).

En la prueba exacta de Fisher, aconsejable cuando exista alguna celda con frecuencias menores que 5, seleccionamos el estadístico calculado para la hipótesis bilateral (*two tailed*), salvo que supongamos a priori, antes de ver nuestros datos, que las personas que no fuman tendrán una mejor respuesta al fármaco, en cuyo caso utilizaríamos la prueba unilateral (*one tailed*). Como la $p < 0{,}001$, de nuevo se rechaza la hipótesis nula de independencia y se concluye que las personas que fuman y las que no fuman tienen una respuesta distinta al fármaco destinado a reducir la hipertensión.

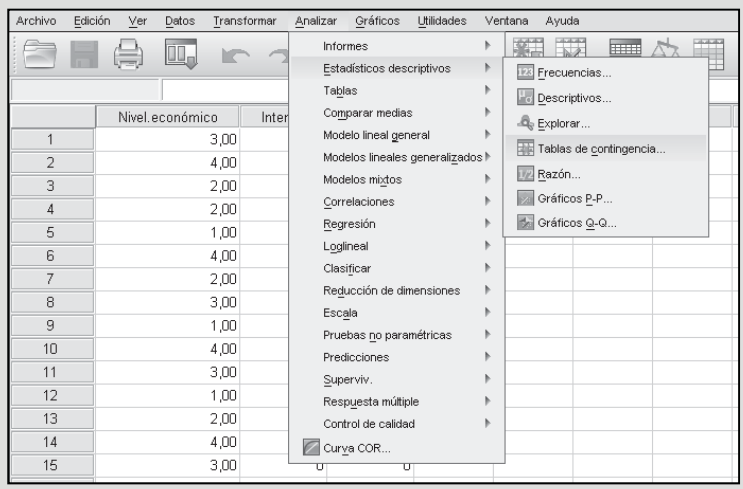
El coeficiente Phi confirma está asociación significativa entre las dos variables con un valor de 0,789, indicando una alta asociación.

CUADRO VIII.2. *(Continuación)*[SPSS]

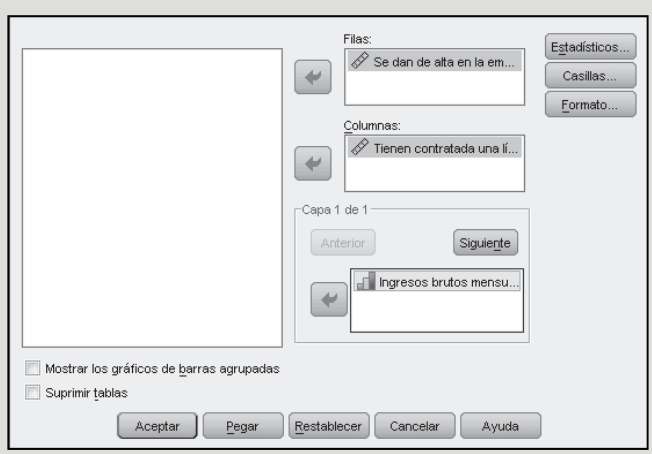
Aplicación con SPSS.

EJEMPLO. Se van a utilizar los mismos datos que se usaron con R, es decir, el objetivo es determinar si el que acepten o no acepten la oferta de la empresa para contratar una conexión de internet de mayor velocidad depende de que ya tengan o no tengan internet y/o del salario bruto mensual. Los datos están en el archivo **Cuadro VIII.2.sav**.

Paso 1. Se selecciona «Analizar» ⇨ «Estadísticos descriptivos» ⇨ «Tablas de contingencia…».

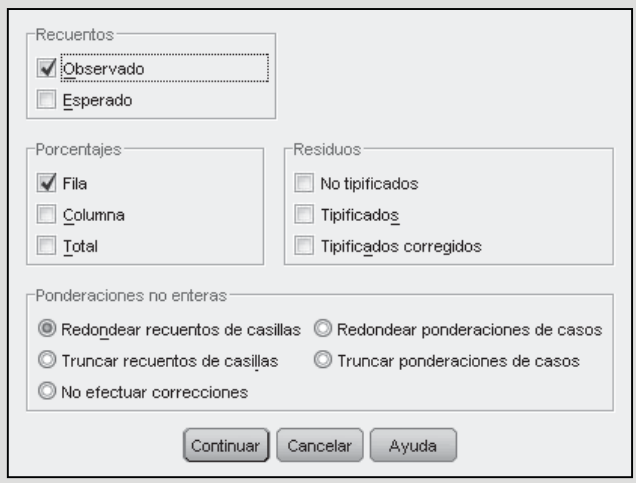


Paso 2. En «Filas» hay que poner la variable dependiente, es decir, si contrata o no la nueva línea de internet con la empresa. En «Columnas» hay que poner la variable independiente, es decir, si ya tiene internet. Por último, como queremos saber si la contratación de la nueva línea de internet depende del ingreso bruto mensual, en «Capa» se pone la variable que informa sobre el nivel económico.

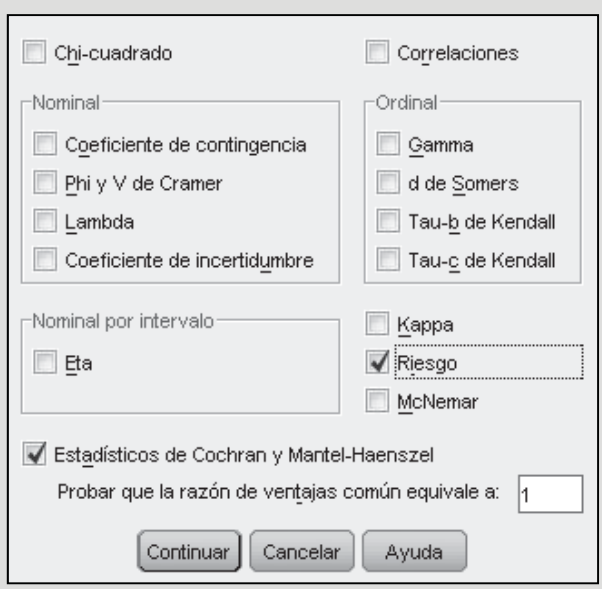


CUADRO VIII.2. *(Continuación)*[SPSS]

Paso 3. En la ventana anterior en «Casillas…» hay que marcar «Observado» y «Fila» para que en la tabla de contingencia, además de los valores absolutos, también aparezca la información de los porcentajes.



Paso 4. Como es una tabla *K*x2x2, en vez de la prueba de Fisher es mejor usar las pruebas de Cochran y Mantel-Haenszel, y como estimación de la intensidad de la asociación, el riesgo relativo y la razón de ventajas (*odds ratio*). En la ventana del paso 2, en «Estatísticos…» aparece la siguiente ventana donde se selecciona «Riesgo» y «Estadísticos de Cochran y Mantel-Haenszel».



CUADRO VIII.2. *(Continuación)*[SPSS]

Paso 5. La tabla de contingencia con los resultados absolutos y los porcentajes se muestra a continuación.

Ingresos brutos mensuales				Tienen contratada una línea de internet		Total
				Si	No	
Menos de 1000 €	Se dan de alta en la empresa	Si	Recuento	663	311	974
			% dentro de Se dan de alta en la empresa	68,1%	31,9%	100,0%
		No	Recuento	90	110	200
			% dentro de Se dan de alta en la empresa	45,0%	55,0%	100,0%
	Total		Recuento	753	421	1174
			% dentro de Se dan de alta en la empresa	64,1%	35,9%	100,0%
1000 a 1499 €	Se dan de alta en la empresa	Si	Recuento	1366	720	2086
			% dentro de Se dan de alta en la empresa	65,5%	34,5%	100,0%
		No	Recuento	146	156	302
			% dentro de Se dan de alta en la empresa	48,3%	51,7%	100,0%
	Total		Recuento	1512	876	2388
			% dentro de Se dan de alta en la empresa	63,3%	36,7%	100,0%
1500 a 2000 €	Se dan de alta en la empresa	Si	Recuento	598	420	1018
			% dentro de Se dan de alta en la empresa	58,7%	41,3%	100,0%
		No	Recuento	36	66	102
			% dentro de Se dan de alta en la empresa	35,3%	64,7%	100,0%
	Total		Recuento	634	486	1120
			% dentro de Se dan de alta en la empresa	56,6%	43,4%	100,0%
más de 2000 €	Se dan de alta en la empresa	Si	Recuento	688	872	1560
			% dentro de Se dan de alta en la empresa	44,1%	55,9%	100,0%
		No	Recuento	45	113	158
			% dentro de Se dan de alta en la empresa	28,5%	71,5%	100,0%
	Total		Recuento	733	985	1718
			% dentro de Se dan de alta en la empresa	42,7%	57,3%	100,0%

CUADRO VIII.2. *(Continuación)*[SPSS]

Paso 6. Las pruebas de Cochran y Mantel-Haenszel son significativas (p < 0,001). Por tanto, se rechaza la hipótesis nula de independencia y concluimos que existe una relación significativa entre que una persona acepte la oferta de contratar una línea de internet de más velocidad y que ya tenga internet en casa. Es importante mencionar que estas pruebas ya tienen en cuenta el efecto de salario bruto mensual de la persona, es decir, no solo considera una tabla 2x2, sino todas las tablas de todos los estratos (todos los niveles de salario bruto).

Pruebas de independencia condicional

	Chi-cuadrado	gl	Sig. asintótica (bilateral)
De Cochran	103,579	1	,000
Mantel-Haenszel	102,705	1	,000

Paso 7. En la tabla que se muestra en la siguiente página aparecen los valores del predominio o ventaja de que la persona acepte la oferta de la empresa de contratar la línea de internet (cociente entre la probabilidad de que acepte y la probabilidad de que no acepte).

Por ejemplo, el predominio de que una persona que gana menos de 1.000€ y que ya tiene contratada una línea de internet, acepte la nueva línea de internet con más velocidad es 1,513. El cálculo es a partir de la tabla de contingencia: 68,1%/45%. La ventaja o predominio de que una persona que no tiene contratada una línea de internet acepte la nueva línea es 0,581 (31,9%/55%). Por lo tanto, una persona que gana menos de 1.000€ al mes y ya tiene internet tiene 1,513 veces más probabilidad de aceptar la oferta de la empresa que de no aceptarla.

El predominio de una persona que gana entre 1.000 y 1499€ y que ya tiene contratada una línea de internet, de que acepte la nueva línea de internet con más velocidad es 1,355. El cálculo es a partir de la tabla de contingencia: 65,5%/48,3%.

Estos resultados se utilizan para calcular la razón de ventajas OR (*odds ratio*) como se puede ver en la tabla de la siguiente página. Para una persona que gana menos de 1.000€ al mes, la razón de ventajas, que cuantifica el efecto sobre el alta en la empresa de tener ya una línea de internet, es 2,606= 1,513/0,581. Para una persona que gana entre 1.000 y 1499 € es 2,027 =1,355/0,668.

CUADRO VIII.2. *(Continuación)* *SPSS*

Estimación de riesgo

Ingresos brutos mensuales		Valor	Intervalo de confianza al 95%	
			Inferior	Superior
Menos de 1000 €	Razón de las ventajas para Se dan de alta en la empresa (Si / No)	2,606	1,912	3,551
	Para la cohorte Tienen contratada una línea de internet = Si	1,513	1,290	1,774
	Para la cohorte Tienen contratada una línea de internet = No	,581	,497	,678
	N de casos válidos	1174		
1000 a 1499 €	Razón de las ventajas para Se dan de alta en la empresa (Si / No)	2,027	1,590	2,585
	Para la cohorte Tienen contratada una línea de internet = Si	1,355	1,201	1,528
	Para la cohorte Tienen contratada una línea de internet = No	,668	,590	,756
	N de casos válidos	2388		
1500 a 2000 €	Razón de las ventajas para Se dan de alta en la empresa (Si / No)	2,610	1,707	3,992
	Para la cohorte Tienen contratada una línea de internet = Si	1,664	1,273	2,175
	Para la cohorte Tienen contratada una línea de internet = No	,638	,543	,749
	N de casos válidos	1120		
más de 2000 €	Razón de las ventajas para Se dan de alta en la empresa (Si / No)	1,981	1,383	2,839
	Para la cohorte Tienen contratada una línea de internet = Si	1,548	1,202	1,995
	Para la cohorte Tienen contratada una línea de internet = No	,782	,702	,871
	N de casos válidos	1718		

Paso 8. La razón de ventajas es útil para comparar en los distintos estratos, es decir, en nuestro caso en función del dinero bruto mensual que gana la persona, la intensidad de la asociación que existe entre que contrate la nueva línea y que ya tuviese internet.

CUADRO VIII.2. *(Continuación)*[SPSS]

Tanto en la prueba de Breslow-Day como la de Tarone la probabilidad es mayor de 0,05 (p = 0,474 en ambos casos). Como son pruebas de homogeneidad, se acepta la hipótesis nula de que no hay diferencias significativas en la razón de ventajas entre los distintos estratos. Es decir, la intensidad de la asociación entre las dos variables (que acepte la oferta de internet de mayor velocidad y que ya tenga internet) no varía en función del salario bruto mensual.

Pruebas de homogeneidad de la razón de las ventajas

	Chi-cuadrado	gl	Sig. asintótica (bilateral)
Breslow-Day	2,509	3	,474
De Tarone	2,508	3	,474

Como no existen diferencias significativas entre las diferentes categorías de salario bruto mensual, es posible dar un valor de la razón de ventajas que incluye todos los estratos y que es 2,225.

Estimación de la razón de las ventajas común de Mantel-Haenszel

Estimación			2,225
ln(estimación)			,800
Error típ. de ln(estimación)			,080
Sig. asintótica (bilateral)			,000
Intervalo de confianza asintótico al 95%	Razón de ventajas común	Límite inferior	1,902
		Límite superior	2,603
	ln(Razón de ventajas común)	Límite inferior	,643
		Límite superior	,957

La estimación de la razón de las ventajas común de Mantel-Haenszel se distribuye de manera asintóticamente normal bajo el supuesto de razón de las ventajas común igual a 1,000. Lo mismo ocurre con el log natural de la estimación.

Una razón de ventajas igual a 1 indicaría que el predominio de altas es el mismo entre aquellos que ya tienen línea de internet y aquellos que no la tienen. El valor p < 0,001 (designado por SPSS significación asintótica bilateral) indica que podemos descartar esa hipótesis y, por lo tanto, la razón de ventajas es significativamente distinta de uno. Su valor estimado 2,225 nos permite decir que el predominio de nuevas altas es algo más del doble entre aquellos que ya tienen contratada una línea de internet en relación con los que no la tienen, por lo que la empresa debería centrar sus esfuerzos en ofertar la línea de internet de alta velocidad a aquellas personas que ya tienen internet en casa, sin importar el salario bruto mensual de la persona.

Capítulo IX

CONTRASTES DE HOMOGENEIDAD EN VARIABLES CUANTITATIVAS

En el capítulo anterior se analizaron los contrastes de homogeneidad e independencia en variables cualitativas. En este capítulo se explicarán los métodos para hacerlo con variables cuantitativas.

Para el contraste de homogeneidad en el caso de variables cuantitativas contamos con dos tipos de análisis, los *paramétricos,* que proceden de un modelo que obliga a cumplir ciertos supuestos acerca de los parámetros y la distribución de probabilidad de la población de la que se extrajo la muestra, y los *no paramétricos,* que son menos restrictivos para poder ser aplicados. Se debe tener en cuenta que la estadística paramétrica es más precisa, pero las posibilidades de ser aplicada, como veremos más adelante, están limitadas.

IX.1. Pruebas paramétricas

IX.1.1. REQUISITOS

Antes de utilizar las pruebas paramétricas, es indispensable comprobar ciertos requisitos (supuestos) para su aplicación. Los más importantes a tener en cuenta son:

1. Distribución Normal de las poblaciones.
2. Igualdad de varianzas.

Para la comprobación de la hipótesis de Normalidad hay que realizar un contraste de bondad de ajuste y podemos emplear la prueba de Kolmogorov–Smirnov o la de Shapiro-Wilk explicadas en el Capítulo VI (véase Cuadro VI.2). La prueba de Normalidad se debe hacer sobre los residuos o diferencias entre los valores observados de la variable dependiente y los valores que predice el modelo una vez estimados los parámetros. Si los residuos son Normales, también la variable dependiente tendrá distribución Normal en cada nivel de cada factor.

La hipótesis de igualdad de varianzas se verifica comúnmente como se detalló en el Capítulo IV (véase Cuadro IV.4) o como se explica en los ejemplos que veremos a continuación.

Si nuestros datos no cumplen el supuesto de distribución Normal se puede intentar transformar la variable para que adopte una distribución Normal, como veremos en el siguiente apartado. En general, la transformación que "normaliza" los datos también consigue una igualdad de varianzas. En el caso de que no se consiguiera la distribución Normal y la homogeneidad de varianzas, ni siquiera después de transformar los datos, la mejor alternativa sería utilizar un método no paramétrico (véase apartado IX.2), aunque a veces algunos tests paramétricos se aplican no suponiendo la igualdad de varianzas, como por ejemplo el *t*-test modificado por Welch-Satterthwaite (véase apartado IX.1.3.1).

IX.1.2. TRANSFORMACIONES

Si decidimos transformar la variable, tendremos varias posibilidades según su tipo de distribución (asimétrica positiva o negativa, ver Capítulo III). La bibliografía nos habla de la llamada *escalera de las transformaciones* de Tukey, la cual muestra el tipo de transformación recomendada según sea la intensidad de la asimetría o la dirección en la que van los casos extremos (Sánchez 1999). La Figura IX.1 es una modificación del gráfico realizado por Erickson & Nosanchuk (1977).

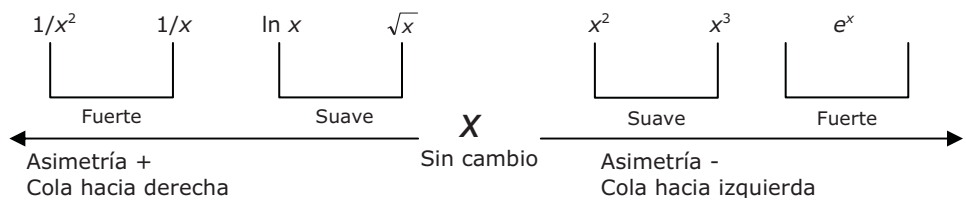


Figura IX.1. Escalera de transformaciones adaptada de Erickson & Nosanchuk (1977).

Para elegir la transformación adecuada hay que tener en cuenta la asimetría y la curtosis. Como la gráfica solo muestra la asimetría, la influencia de la curtosis puede dar lugar a que valores inferiores a la media requieran una transformación de raíz cuadrada y valores superiores a la media requieran elevar al cuadrado. Se debe elegir la transformación que haga más próximos a cero simultáneamente ambos coeficientes.

IX.1.3. *t*-TEST

IX.1.3.1. Muestras independientes

El *t*-test es el método más común para evaluar las diferencias entre las medias de dos grupos independientes, por ejemplo dos grupos de peces sometidos a dietas diferentes. Para esta prueba, idealmente los sujetos deben asignarse aleatoriamente a dos grupos, de forma que cualquier diferencia en la respuesta sea debida al tratamiento (o falta de tratamiento) y no a otros factores. Al aplicar este análisis debemos asegurarnos de que las diferencias en otros factores no enmascaren o resalten una diferencia significativa entre las medias.

La hipótesis nula (H_o) con la que comúnmente se trabaja es que las medias son iguales. En general no se exige que el número de observaciones en ambos grupos sea igual. Si la H_o es cierta y suponemos una igualdad de varianzas, el estadístico (*t*-test) seguirá una distribución *t* de Student con $n_x + n_y - 2$ grados de libertad. Existe también la opción de utilizar el estadístico *t* de Student asumiendo una no igualdad de varianzas, al trabajar con los grados de libertad de Welch-Satttherthwaite.

En el Cuadro IX.1 se muestra un ejemplo del *t*-test para muestras independientes

CUADRO IX.1. *t*-test para muestras independientes

Aplicación con R.

EJEMPLO. En el archivo **Cuadro IX.1.V.csv** hay datos de peso y altura de niños y niñas entre los 2 y 4 años en Italia y España. El objetivo del trabajo es determinar si existen diferencias significativas en el peso de los niños de 2 y 3 años entre ambos países.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro IX.1.R**, que se muestra en la siguiente ventana. Sólo es necesario especificar lo que se solicita en la sección modificable por el usuario. Para que este *script* funcione correctamente es necesario tener instalados los paquetes "mvtnorm", "VGAM", "splines", "stats4", "lawstat" y "e1071", o realizar la instalación como se explica en el Apéndice II.

```
#Cuadro IX.1
#Test t para muestras independientes.

################################################
#Sección modificable por el usuario
################################################

#Lectura de datos.
datos<-read.csv2("Cuadro IX.1.V.csv",header=TRUE,encoding="latin1")

#Selección variable dependiente. En caso de que se deseen utilizar valores
varInteres<-c("Peso")
#varInteres<-c("Altura")

#Selección de un tipo de transformación para las variables de interés. Si r
varInteres1X2<-NULL
varInteres1X<-NULL
varInteresLN<-NULL
varInteresLOG10<-NULL
varInteresSQR<-NULL
varInteresX2<-NULL
varInteresX3<-NULL
varInteresEX<-NULL
varInteresASN<-NULL

#varInteres1X2<-c(1,1)
#varInteres1X<-c(1,1)
#varInteresLN<-c(1,1)
#varInteresLOG10<-c(1,1)
#varInteresSQR<-c(1,1)
#varInteresX2<-c(1,1)
#varInteresX3<-c(1,1)
#varInteresEX<-c(1,1)
#varInteresASN<-c(1,1)
```

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro IX.1.V.csv**.

datos<-read.csv2("Cuadro IX.1.V.csv",header=TRUE,encoding="latin1")

CUADRO IX.1. *(Continuación)*[R]

Paso 2. Se selecciona la variable dependiente, en este caso el peso.

```
varInteres<-c("Peso")
#varInteres<-c("Altura")
```

Paso 3. En caso de que sea necesario es posible transformar la variable dependiente, para ello simplemente hay que quitar el símbolo # en el tipo de transformación que se quiera realizar. En primer lugar se hace el análisis sin transformar los datos, y en el caso de no cumplirse la Normalidad y/o homogeneidad de varianzas, entonces se realizaría el análisis aplicando la transformación más adecuada.

```
#varInteres1X2<-c(1,1)
#varInteres1X<-c(1,1)
#varInteresLN<-c(1,1)
#varInteresLOG10<-c(1,1)
#varInteresSQR<-c(1,1)
#varInteresX2<-c(1,1)
#varInteresX3<-c(1,1)
#varInteresEX<-c(1,1)
#varInteresASN<-c(1,1)
```

Paso 4. A continuación se indica cual es el factor en «varPoblacion» y se seleccionan los dos grupos que se van a comparar. En el caso de que sólo existan dos grupos, es posible poner la opción «NULL» sin necesidad de especificar los dos grupos a comparar. En el ejemplo, como se quiere determinar las diferencias entre España y Italia, se ponen esos dos grupos, aunque también se podría haber puesto «NULL», ya que ese factor tiene solo dos clases.

```
varPoblacion<-c("País")
varComparar1<-c("España")
varComparar2<-c("Italia")
#varComparar1<-NULL
```

Paso 5. También es posible seleccionar grupos en dos variables diferentes (un máximo de dos grupos en cada variable). En el ejemplo como se quiere comparar solo los niños, se selecciona este grupo en la variable «Sexo» y las edades de 2 y 3 años en la variable «Edad».

```
varSeleccion1<-list(c("Sexo","Niño"))
#varSeleccion1<-NULL
varSeleccion2<-list(c("Edad","2","3"))
#varSeleccion2<-NULL
```

Paso 6. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
ArchivodeSalida<-"Salida Cuadro IX.1.txt"
```

CUADRO IX.1. *(Continuación)*[R]

Paso 7. El test de Shapiro-Wilk muestra que las dos poblaciones de niños en España (p = 0,704) e Italia (p = 0,933) tienen distribución Normal. Además se cumple el requisito de homogeneidad de varianzas, tanto considerando la media (p = 0,864) como la mediana (p = 0,916). Por lo tanto, no es necesario realizar de nuevo el análisis transformando la variable dependiente. En el caso de que no se cumpliera alguno de los dos requisitos, también sale información sobre el asimetría y la curtosis de cada una de las poblaciones, lo cual ayudaría a elegir mejor la transformación a realizar.

```
[[1]]
datos3[, "poblaciones"]: España

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9676, p-value = 0.7042

------------------------------------------------------------
datos3[, "poblaciones"]: Italia

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9799, p-value = 0.9333


[[2]]

        classical Levene's test based on the absolute deviations from the mean
        ( none not applied because the location is not set to median )

data:  datos3$valores
Test Statistic = 0.0295, p-value = 0.8645


[[3]]

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  datos3$valores
Test Statistic = 0.0112, p-value = 0.9164


[[4]]
                                España                Italia
            "Sesgo" "0.346732020676607" "0.241321686507485"

[[5]]
                                   España                 Italia
          "Curtosis" "-0.546546454680671" "-0.543244470678692"

[[6]]

        Welch Two Sample t-test

data:  datos3$valores by datos3$poblaciones
t = 0.4759, df = 37.962, p-value = 0.6369
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6344797  1.0244797
sample estimates:
mean in group España mean in group Italia
              12.935                12.740
```

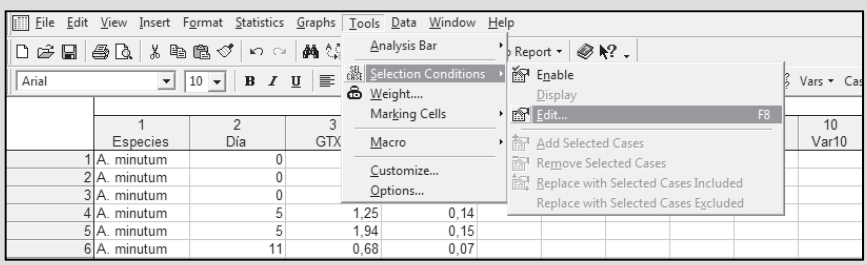
Por último, el test de la t muestra que no existen diferencias significativas en el peso de los niños de 2 y 3 años entre España e Italia (t =0,47, df = 38, p = 0,637), con valores medios de 12,93 y 12,74 kgs, respectivamente.

CUADRO IX.1. *(Continuación)*[STATISTICA]

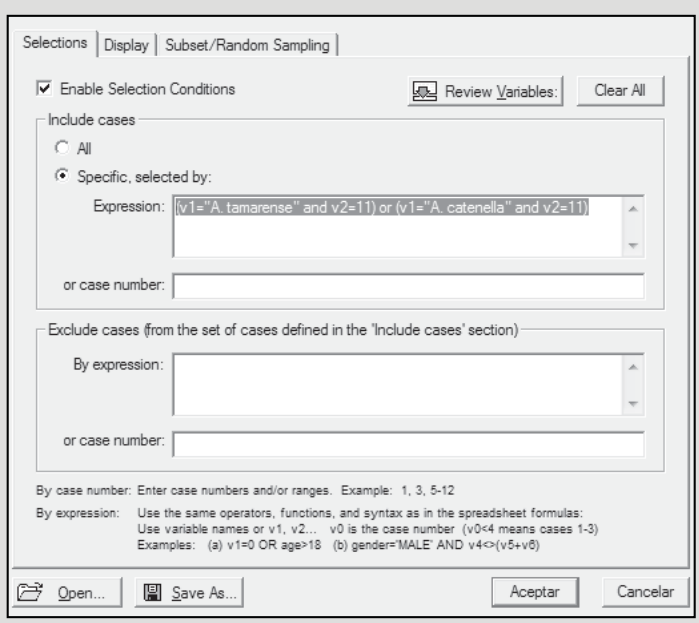
Aplicación con STATISTICA.

EJEMPLO. En el archivo **Cuadro IX.1.sta** hay datos de la cantidad de dos toxinas, GTX4 y GTX1 (en pg célula^{-1}), en tres especies de fitoplancton tóxico del género *Alexandrium* a lo largo del tiempo: *A. minutum, A. catenella* y *A. tamarense*. El objetivo es determinar si existen diferencias significativas en la cantidad de la toxina GTX4 entre las especies *A. tamarense* y *A. catenella* sólo en el día 11 del experimento.

Paso 1. Debido a que existen varias especies y varios días de experimento, es necesario primero definir las especies y los días con los que se va a trabajar. Para seleccionar los casos vamos a «Tools» ⇨ «Selection Conditions» ⇨ «Edit…».

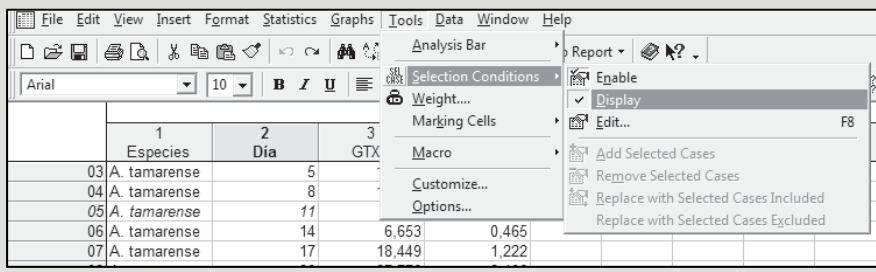


Paso 2. En la pestaña «Selections» se marca «Enable Selection Conditions» ⇨ «Specific, selected by:» donde se selecciona la variable o variables a analizar, usando OR y AND si se quieren combinar varias condiciones. En este caso se selecciona «(v1="A. tamarense" and v2=11) or (v1="A. catenella" and v2=11)», es decir, las especies *A. tamarense* y *A. catenella* en el día 11 del experimento.

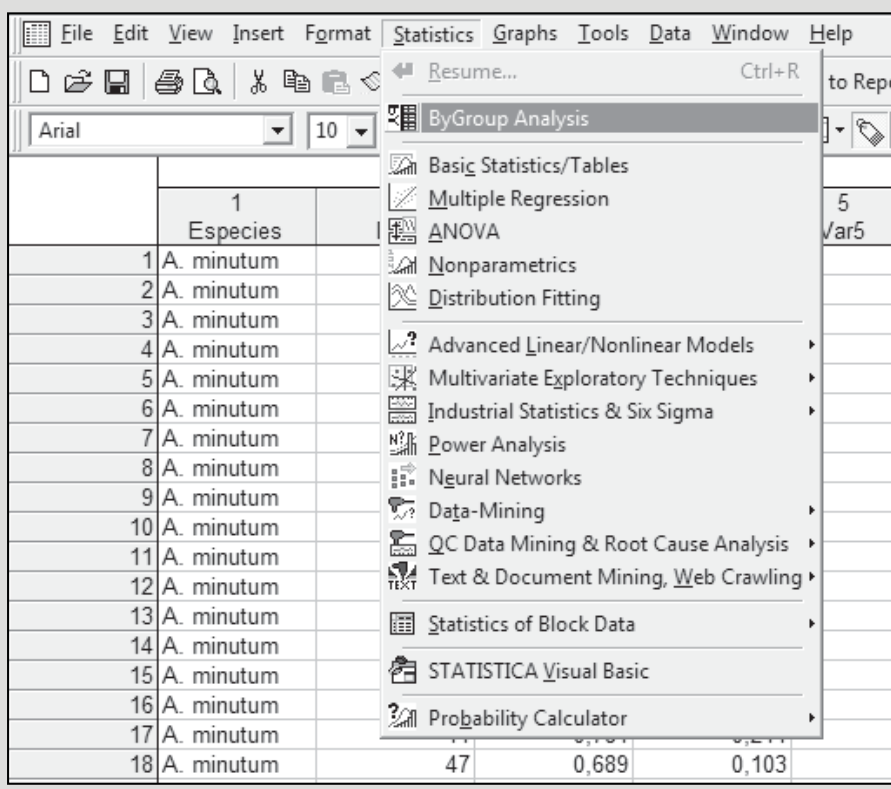

Selections
Display
Subset/Random Sampling
Enable Selection Conditions
Review Variables:
Clear All
Include cases
All
Specific, selected by:
Expression: (v1="A. tamarense" and v2=11) or (v1="A. catenella" and v2=11)
or case number:
Exclude cases (from the set of cases defined in the 'Include cases' section)
By expression:
or case number:
By case number: Enter case numbers and/or ranges. Example: 1, 3, 5-12
By expression: Use the same operators, functions, and syntax as in the spreadsheet formulas:
Use variable names or v1, v2... v0 is the case number (v0<4 means cases 1-3)
Examples: (a) v1=0 OR age>18 (b) gender='MALE' AND v4<>(v5+v6)
Open...
Save As...
Aceptar
Cancelar

CUADRO IX.1. *(Continuación)*[STATISTICA]

Paso 3. Para resaltar los casos seleccionados vamos a «Tools» ⇨ «Selection Conditions» ⇨ «Display». En la tabla de datos aparecerán en cursiva los valores seleccionados. En el supuesto de que no aparecieran en cursiva, eso significaría que la selección en el paso 2 no se hizo de forma correcta.

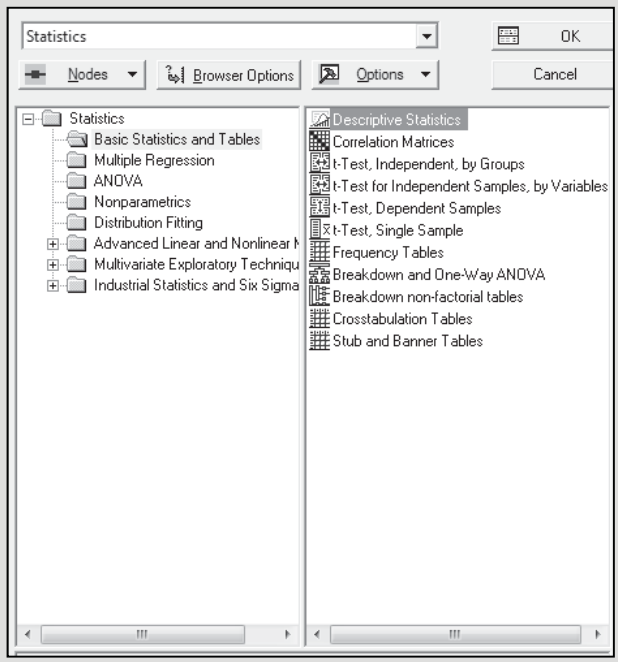


Paso 4. Antes de realizar el *t*-test es necesario comprobar los supuestos de Normalidad y homogeneidad de varianzas. Para testar la Normalidad se puede hacer como se explicó en el Cuadro VI.2 o se selecciona «Statistics» ⇨ «ByGroup Analysis».

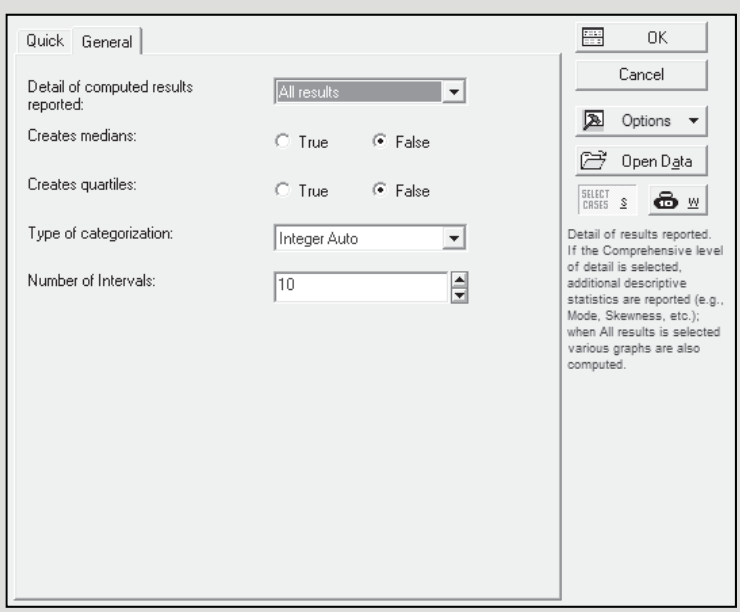


CUADRO IX.1. *(Continuación)* *STATISTICA*

Paso 5. En la ventana que aparece se selecciona «Basic Statistics and Tables» ⇨ «Descriptive Statistics».

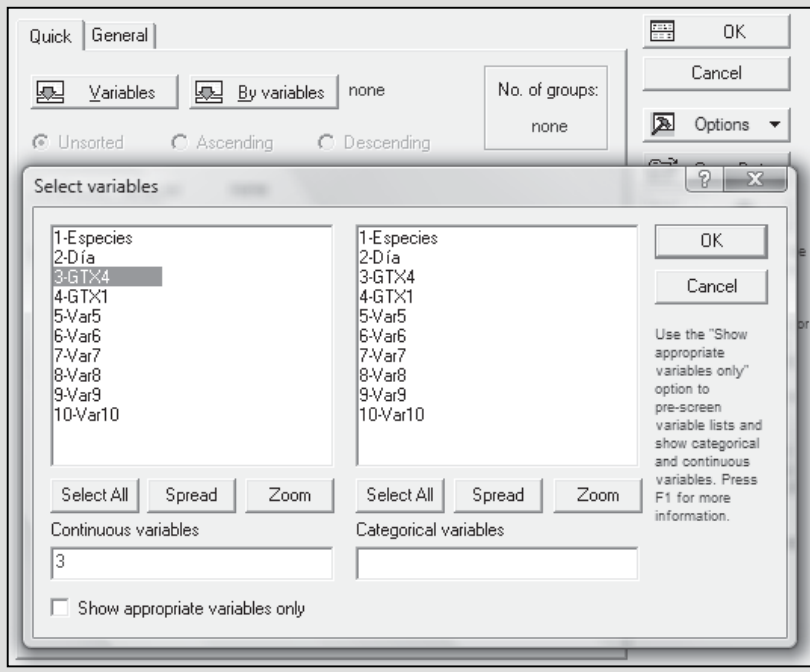


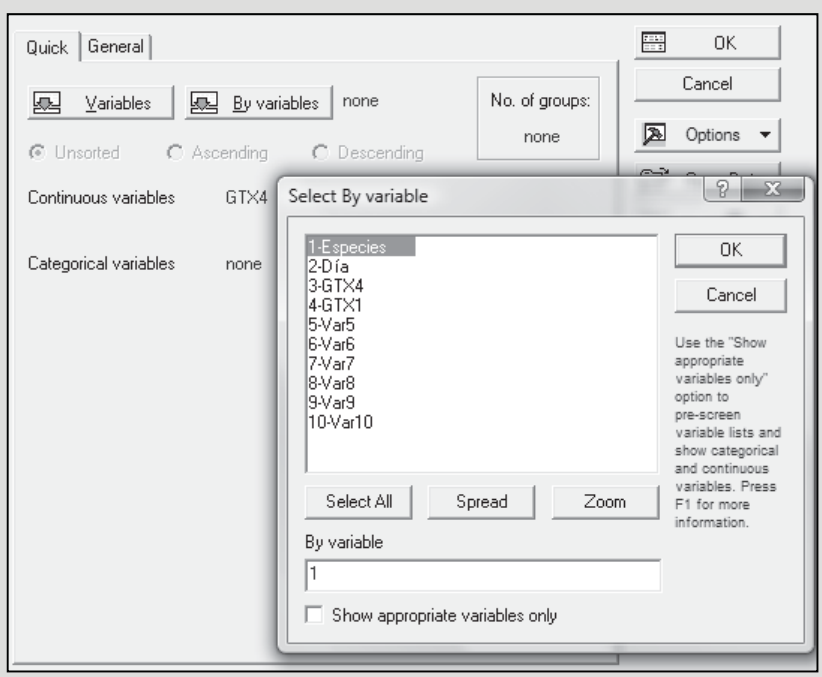
Paso 6. En la siguiente ventana, en la pestaña «General» se selecciona «All results» en «Detail of computed results reported:».



CUADRO IX.1. *(Continuación)*[STATISTICA]

Paso 7. En la pestaña «Quick», en «Variables» se especifica «GTX4» y en «By Variables» se pone «Especies».

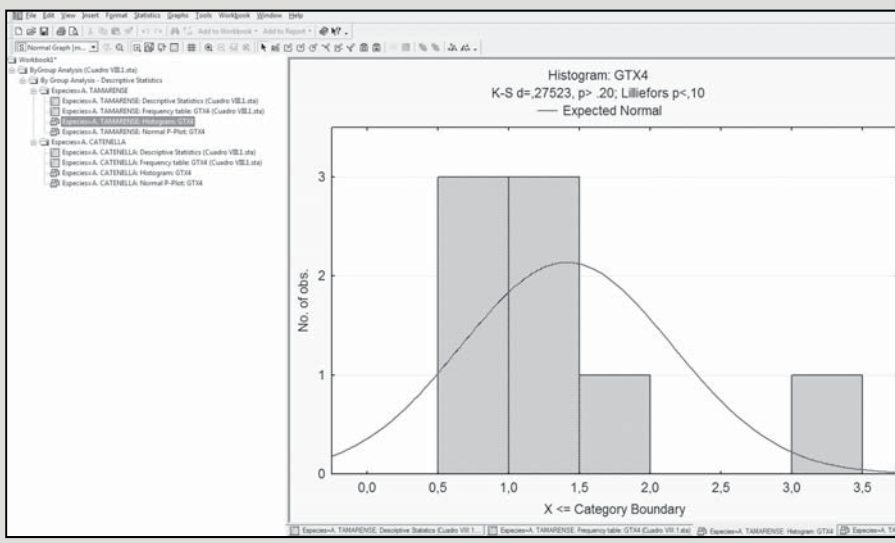




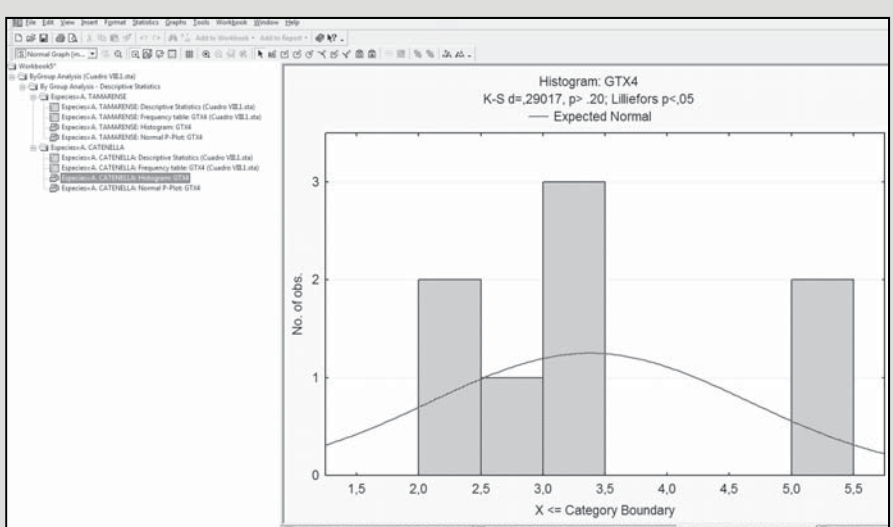
CUADRO IX.1. *(Continuación)*[STATISTICA]

Paso 8. Para poder ver los resultados es necesario pulsar en la ventana de la izquierda en cada una de las combinaciones de las agrupaciones.

En el caso de *A. tamarense* como la probabilidad del test Kolmogorov–Smirnov con la corrección de Lilliefors no es menor de 0,05, se acepta la hipótesis de bondad de ajuste y, por tanto, los datos se ajustan a una distribución Normal.

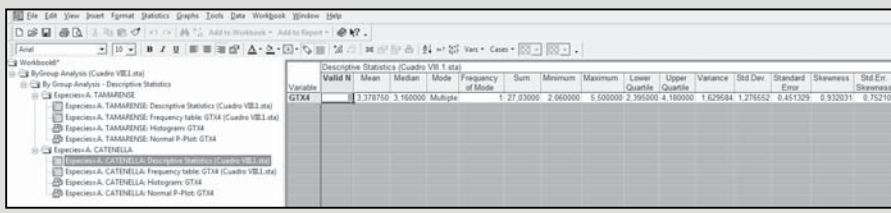


Por el contrario, en el caso de *A. cetenella* como la probabilidad del test Kolmogorov–Smirnov con la corrección de Lilliefors es menor de 0,05, hay que rechazar la hipótesis de bondad de ajuste y, por tanto, los datos no se ajustan a una distribución Normal.

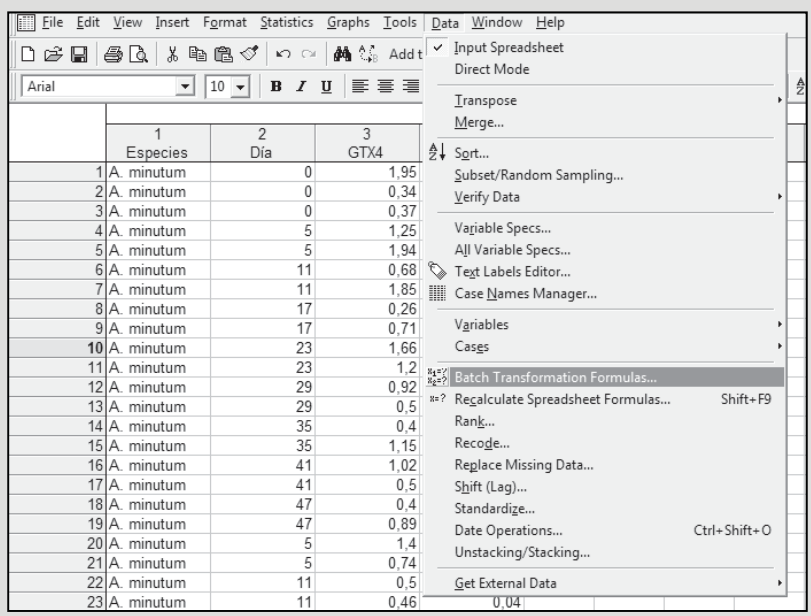


CUADRO IX.1. *(Continuación)*[STATISTICA]

Paso 9. Para poder saber cuál es la transformación más adecuada, en la misma ventana se puede ver el coeficiente de asimetría (*Skewness*) que para *A. catenella* es 0,93. En la Figura IX.1 vemos que para esta transformación lo más adecuado puede ser ln(*x*).

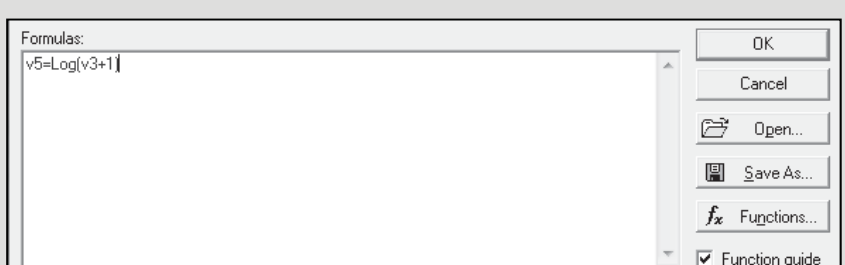


Paso 10. Para poder transformar los datos se selecciona «Data» ⇨ «Batch Transformation Formulas…».



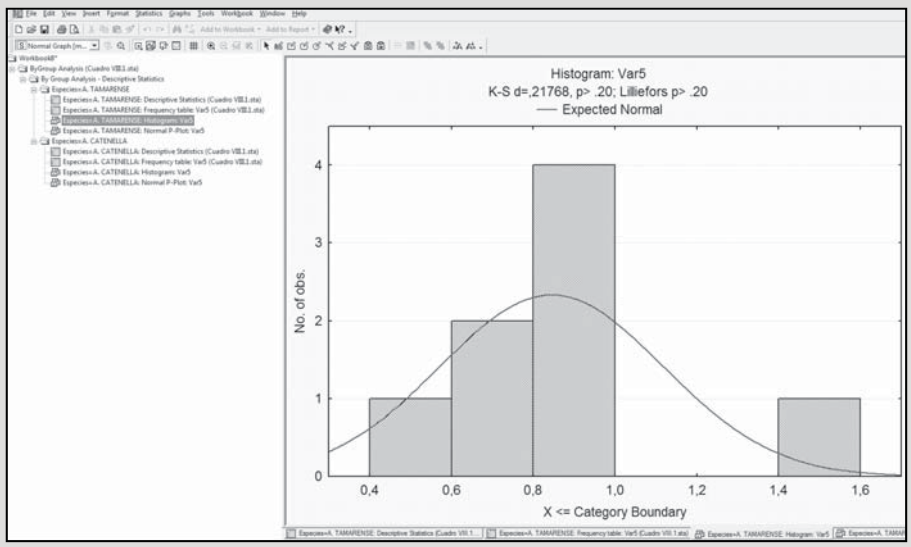
	1 Especies	2 Día	3 GTX4
1	A. minutum	0	1,95
2	A. minutum	0	0,34
3	A. minutum	0	0,37
4	A. minutum	5	1,25
5	A. minutum	5	1,94
6	A. minutum	11	0,68
7	A. minutum	11	1,85
8	A. minutum	17	0,26
9	A. minutum	17	0,71
10	A. minutum	23	1,66
11	A. minutum	23	1,2
12	A. minutum	29	0,92
13	A. minutum	29	0,5
14	A. minutum	35	0,4
15	A. minutum	35	1,15
16	A. minutum	41	1,02
17	A. minutum	41	0,5
18	A. minutum	47	0,4
19	A. minutum	47	0,89
20	A. minutum	5	1,4
21	A. minutum	5	0,74
22	A. minutum	11	0,5
23	A. minutum	11	0,46

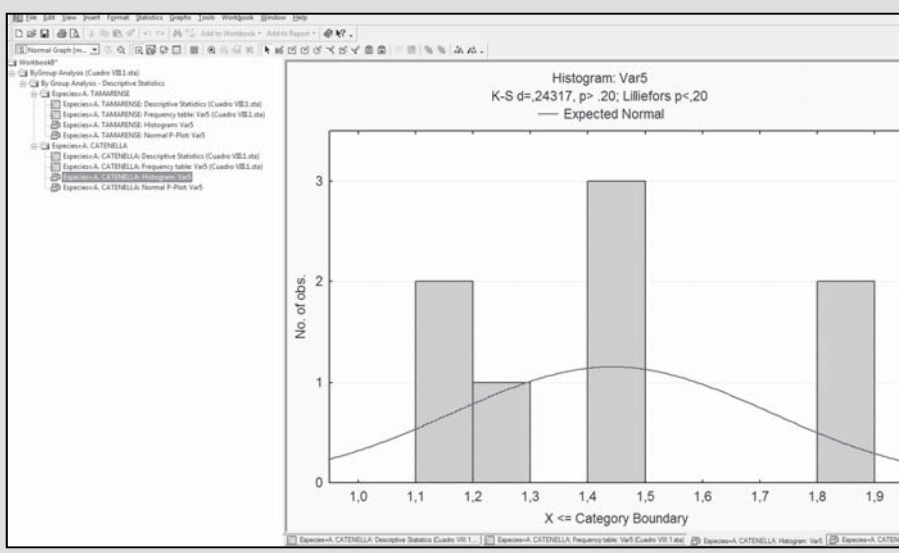
Paso 11. En la ventana siguiente introducimos la transformación «v5=Log(v3+1)». Como puede haber valores cero, es necesario sumar el valor 1 a la variable GTX4 (v3) para poder aplicar el logaritmo.



CUADRO IX.1. *(Continuación)*[STATISTICA]

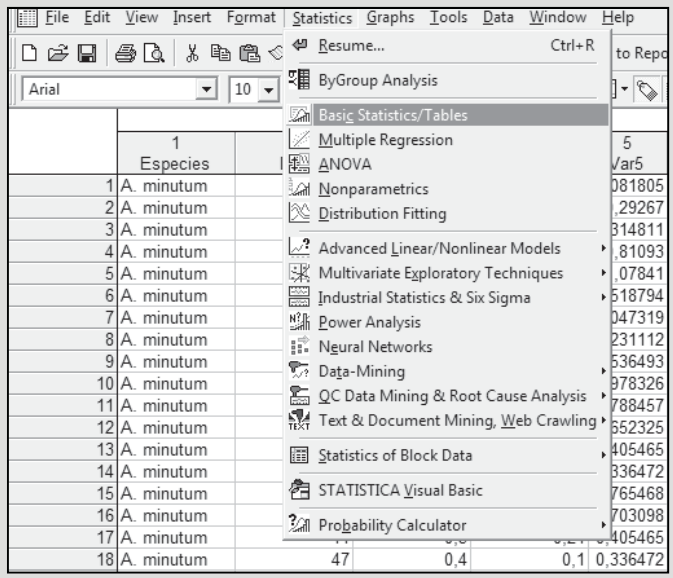
Paso 12. Se repiten los pasos del 4 al 8, pero esta vez en el paso 7, se usa la variable 5 (v5, la variable transformada) en vez de GTX4 (v3). Se observa que tanto para *A. tamarense* como para *A. catenella* $p > 0{,}05$ y, por tanto, en ambos casos se acepta la hipótesis nula de que la distribución es Normal.



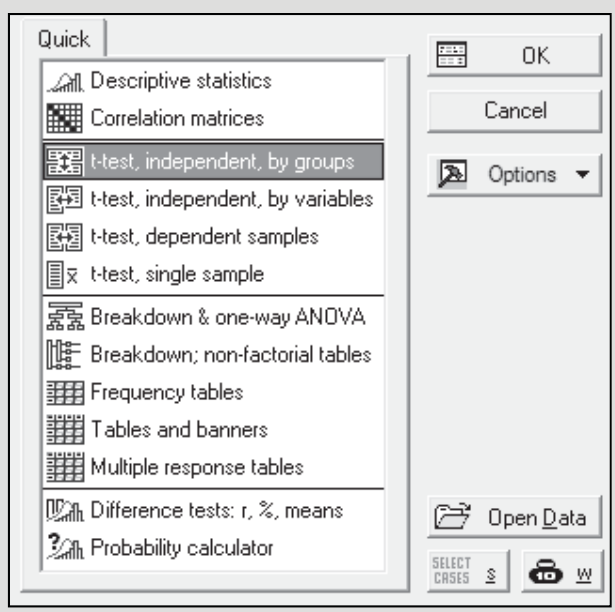


CUADRO IX.1. *(Continuación)*[STATISTICA]

Paso 13. A continuación hay que comprobar si existe homogeneidad de varianzas. Se selecciona «Statistics» ⇨ «Basic Statistics/Tables».

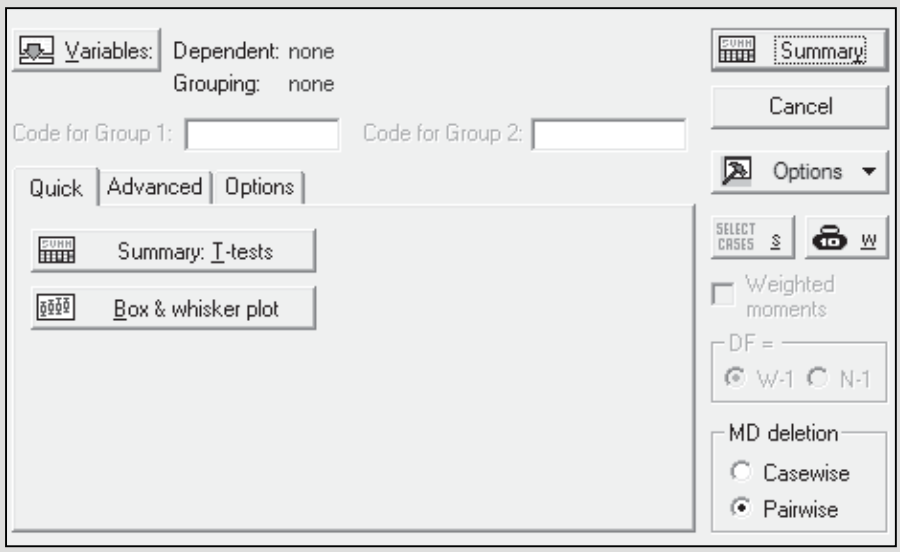


Paso 14. Nos aparece la siguiente ventana donde podemos entrar por dos sitios diferentes dependiendo de cómo tengamos organizados los datos. Si tenemos dos variables en dos columnas, entonces entramos en «t-test, independent, by variables». Si los datos están organizados de tal forma que ambas variables están en una columna y en una segunda hay un código que las identifica, como es el caso en este ejemplo, entonces se entraría en «t-test, independent, by groups».

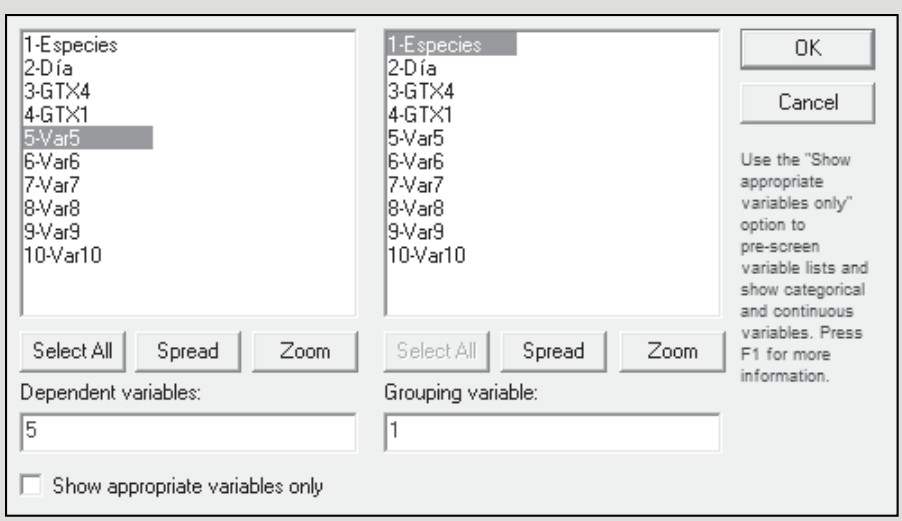


CUADRO IX.1. *(Continuación)*[STATISTICA]

Paso 15. En la ventana siguiente entramos en «Variables» para especificar las variables.

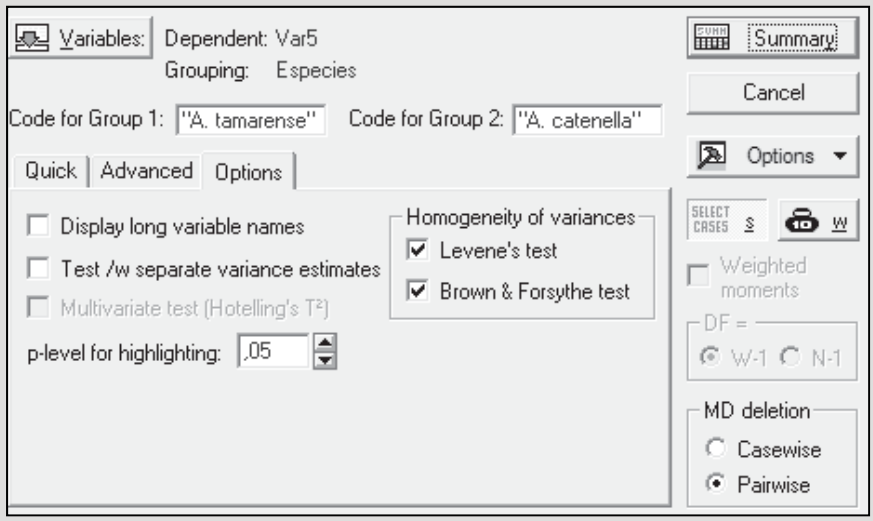


Paso 16. En la ventana que nos sale especificamos la variable que se va a analizar «Dependent variables» (en este caso «Var5») y la que actúa como variable de agrupación «Grouping variable» (en este caso «Especies»).



CUADRO IX.1. *(Continuación)*[STATISTICA]

Paso 17. En la pestaña «Options» marcamos en «Levene´s test» y en «Brown & Forsythe test» y pulsamos «Summary»



Paso 18. Los resultados del test de Levene y del test Brown-Forsythe se muestran en la siguiente tabla. Tanto la probabilidad del Levene test (p = 0,877) como la del test Brown & Forsythe (p = 0,871) son mayores de 0,05, por lo tanto, se acepta la hipótesis nula de que no existen diferencias significativas entre las varianzas.

T-tests; Grouping: Especies (Cuadro VIII.1.sta)
Group 1: A. tamarense
Group 2: A. catenella
Include condition: (v1="A. tamarense" and v2=11) or (v1="A. catenella" and v2=11)

Variable	Std.Dev. A. tamarense	Std.Dev. A. catenella	F-ratio Variances	p Variances	Levene F(1,df)	df Levene	p Levene	Brn-Fors F(1,df)	df Brn-Fors	p Brn-Fors
Var5	0,273078	0,276301	1,023742	0,976103	0,028674	14	0,867957	0,031497	14	0,861678

Paso 19. Como se cumple la Normalidad y homogeneidad de varianzas se puede aplicar el *t*-test, cuyos resultados se muestran también en la misma tabla que los resultados del test de Levene y del test Brown-Forsythe.

T-tests; Grouping: Especies (Cuadro VIII.1.sta)
Group 1: A. tamarense
Group 2: A. catenella
Include condition: (v1="A. tamarense" and v2=11) or (v1="A. catenella" and v2=11)

Variable	Mean A. tamarense	Mean A. catenella	t-value	df	p
Var5	0,844871	1,442191	-4,34898	14	0,000667

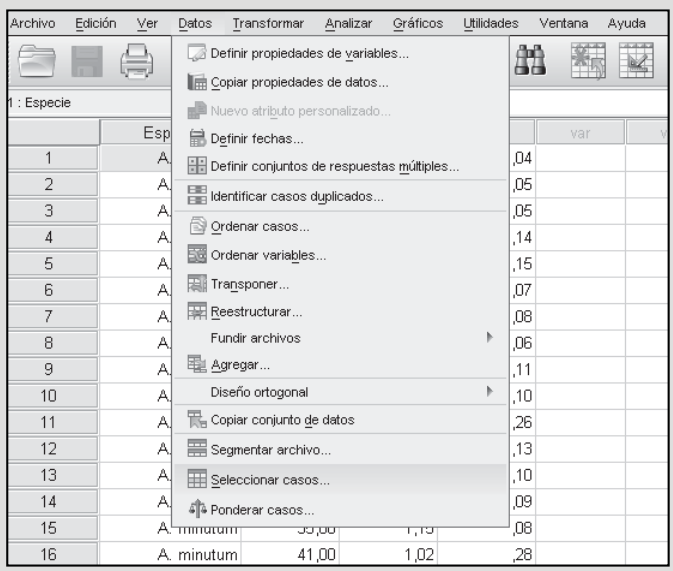
Como $p < 0,05$, se rechaza la hipótesis nula de homogeneidad y, por tanto, existen diferencias significativas en la concentración de toxina GTX4 por célula para el día 11 del experimento entre las especies *A. tamarense* y *A. catenella* (t = -4,34, df = 14, $p < 0,001$).

CUADRO IX.1. *(Continuación)*[SPSS]

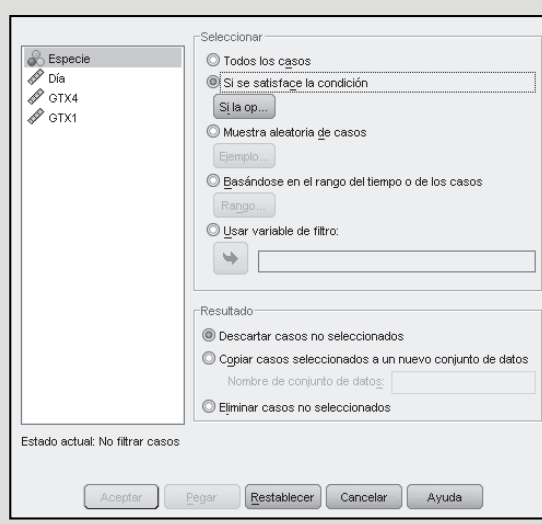
Aplicación con SPSS.

EJEMPLO. Vamos a utilizar los mismos datos que en el caso de la aplicación con STATISTICA, los cuales se encuentran en el archivo **Cuadro IX.1.sav**. En este caso vamos a ver si existen diferencias significativas en la cantidad de la toxina GTX1 entre las especies *A. minutum* y *A. tamarense*, considerando todos los días del experimento.

Paso 1. Como hay tres especies y sólo se quiere trabajar con dos de ellas, en primer lugar hay que seleccionar los grupos a estudiar. Para ello vamos a «Datos» ⇨ «Seleccionar casos…».

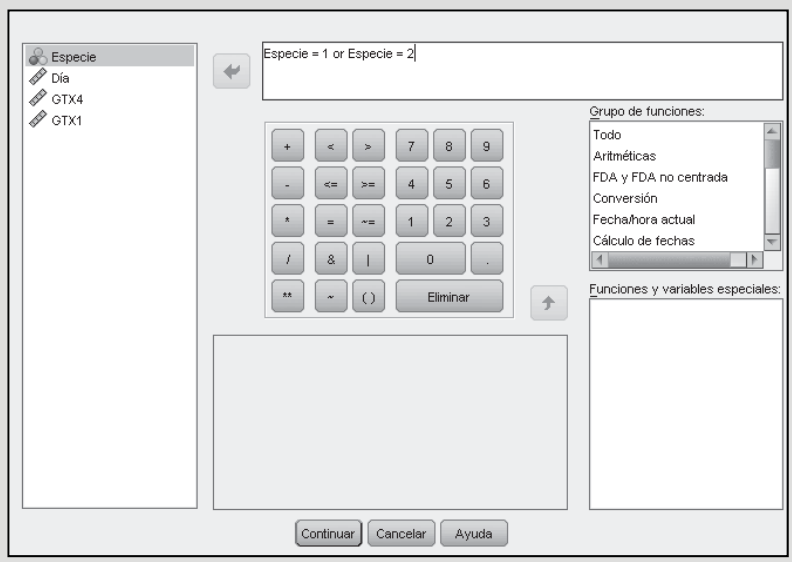


Paso 2. Se selecciona «Si se satisface la condición» y «Si la op…».

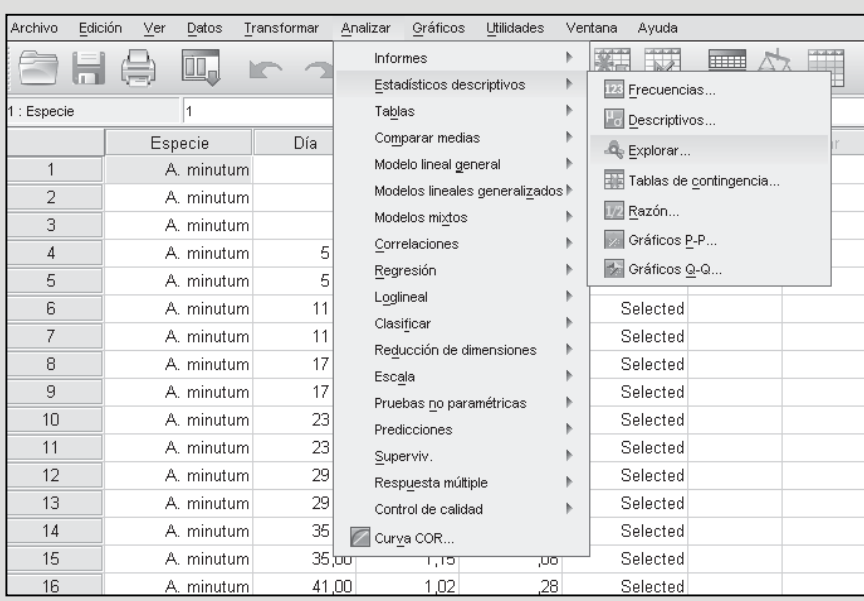


CUADRO IX.1. *(Continuación)*[SPSS]

Paso 3. Se escoge la variable, en este caso «Especie», y se pone la condición «Especie=1 or Especie=2», de tal forma que solo se trabaja con las especies *A. minutum* y *A. tamarense*. Después de pulsar «Continuar», cuando sale la ventana del paso 2 se pulsa «Aceptar».

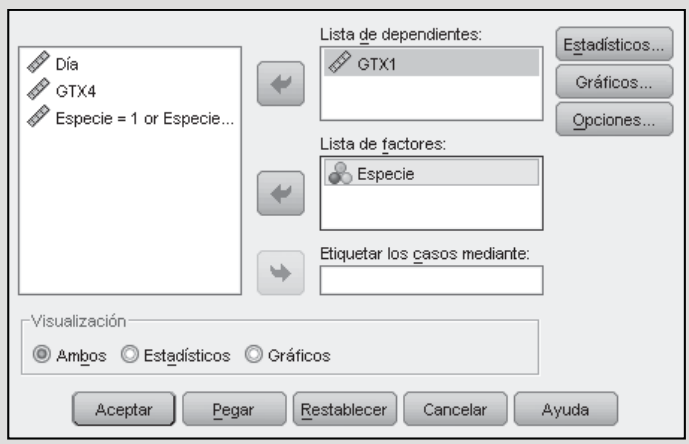


Paso 4. Una vez seleccionados los casos, lo primero que hay que hacer es comprobar si se cumple la distribución Normal y la homogeneidad de varianzas. Se selecciona «Analizar» ⇨ «Estadísticos descriptivos…» ⇨ «Explorar…».

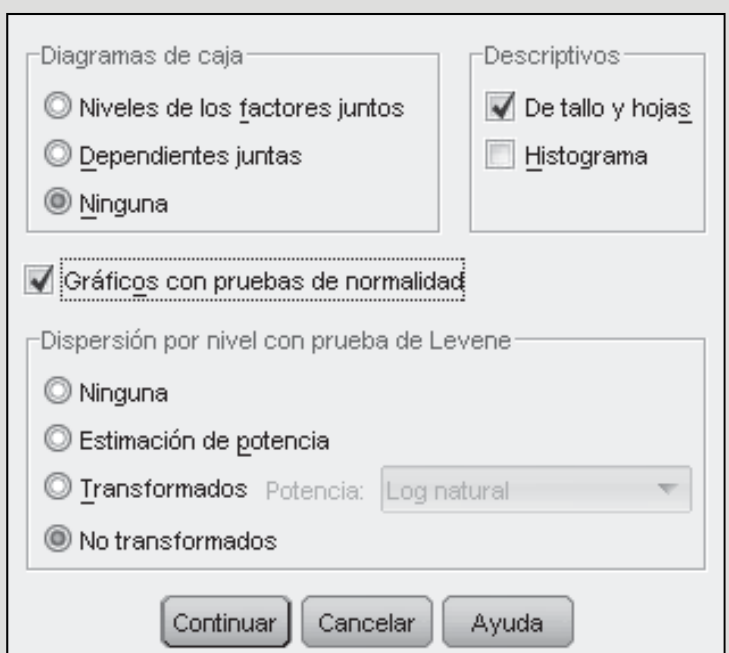


CUADRO IX.1. *(Continuación)*[SPSS]

Paso 5. En la ventana que aparece, en «Lista de dependientes» se incluyen las variables de las que se va a determinar si presentan una distribución Normal, en nuestro ejemplo «GTX1». En «Lista de factores» se incluye «Especie», con lo cual podemos determinar la distribución para las dos especies. En «Visualización» se selecciona «Ambos» y, por último se pulsa el botón «Gráficos».



Paso 6. A continuación en «Diagrama de cajas» se selecciona «Ninguna», se marca «Gráficos con pruebas de normalidad» y, por último, en «Dispersión por nivel con prueba Levene» se marca «No transformados».



CUADRO IX.1. *(Continuación)*[SPSS]

Paso 7. Los estadísticos descriptivos se muestran en la siguiente tabla, donde se observa que para ambas especies la asimetría es positiva.

Descriptivos

	Especie			Estadístico	Error típ.
GTX1	A. minutum	Media		,0951	,00689
		Intervalo de confianza para la media al 95%	Límite inferior	,0813	
			Límite superior	,1088	
		Media recortada al 5%		,0914	
		Mediana		,0900	
		Varianza		,003	
		Desv. típ.		,05636	
		Mínimo		,00	
		Máximo		,28	
		Rango		,28	
		Amplitud intercuartil		,07	
		Asimetría		1,015	,293
		Curtosis		1,791	,578
	A. tamarense	Media		,1806	,01352
		Intervalo de confianza para la media al 95%	Límite inferior	,1536	
			Límite superior	,2076	
		Media recortada al 5%		,1731	
		Mediana		,1700	
		Varianza		,012	
		Desv. típ.		,11067	
		Mínimo		,00	
		Máximo		,54	
		Rango		,54	
		Amplitud intercuartil		,14	
		Asimetría		,983	,293
		Curtosis		1,438	,578

Paso 8. En ambas especies, en el test de Kolmogorov-Smirnov con la corrección de Lilliefors se obtiene una $p > 0,05$, pero como para ambas en el test Shapiro-Wilk la $p = 0,002$, se rechaza la hipótesis nula de que los datos siguen una distribución Normal.

Pruebas de normalidad

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Especie	Estadístico	gl	Sig.	Estadístico	gl	Sig.
GTX1	A. minutum	,107	67	,055	,936	67	,002
	A. tamarense	,097	67	,198	,937	67	,002

a. Corrección de la significación de Lilliefors

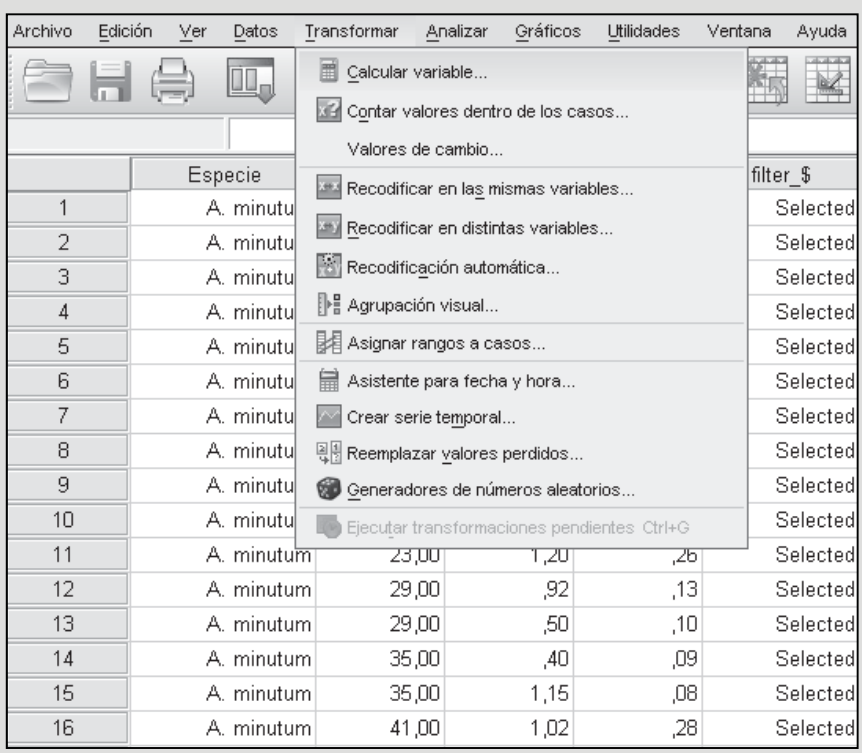
CUADRO IX.1. *(Continuación)*[SPSS]

Paso 9. La prueba de Levene basada en la media muestra que tampoco hay homogeneidad de varianzas ($p < 0{,}001$).

Prueba de homogeneidad de la varianza

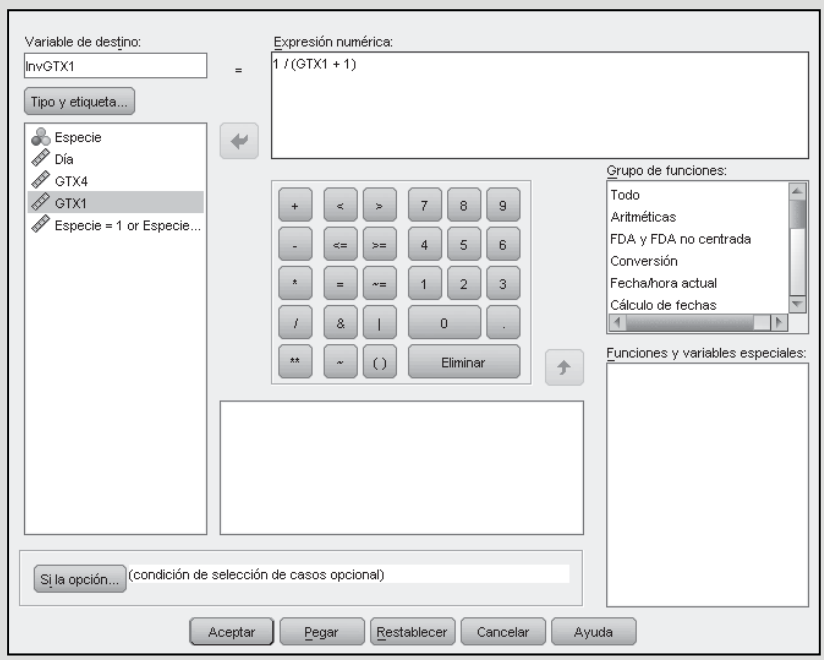
		Estadístico de Levene	gl1	gl2	Sig.
GTX1	Basándose en la media	19,702	1	132	,000
	Basándose en la mediana.	19,265	1	132	,000
	Basándose en la mediana y con gl corregido	19,265	1	100,661	,000
	Basándose en la media recortada	19,471	1	132	,000

Paso 10. Como no se cumple la Normalidad ni la homogeneidad de varianzas es necesario transformar los datos. Para ello vamos a «Transformar» ⇨ «Calcular variable…».


Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana Ayuda
Calcular variable...
Contar valores dentro de los casos...
Valores de cambio...
Recodificar en las mismas variables...
Recodificar en distintas variables...
Recodificación automática...
Agrupación visual...
Asignar rangos a casos...
Asistente para fecha y hora...
Crear serie temporal...
Reemplazar valores perdidos...
Generadores de números aleatorios...
Ejecutar transformaciones pendientes Ctrl+G
Especie
filter_$
1 A. minutu Selected
2 A. minutu Selected
3 A. minutu Selected
4 A. minutu Selected
5 A. minutu Selected
6 A. minutu Selected
7 A. minutu Selected
8 A. minutu Selected
9 A. minutu Selected
10 A. minutu Selected
11 A. minutum 23,00 1,20 ,26 Selected
12 A. minutum 29,00 ,92 ,13 Selected
13 A. minutum 29,00 ,50 ,10 Selected
14 A. minutum 35,00 ,40 ,09 Selected
15 A. minutum 35,00 1,15 ,08 Selected
16 A. minutum 41,00 1,02 ,28 Selected

CUADRO IX.1. *(Continuación)*[SPSS]

Paso 11. Al ser la asimetría alrededor de 1, como se muestra en la Figura IX.1, la transformación que debería ser la más adecuada es ln(x), pero como además no hay homogeneidad de varianzas, es mejor probar la siguiente transformación 1/x. Por ello en «Variable de destino:» creamos una nueva variable «InvGTX1» y en «Expresión numérica» introducimos la transformación «1/(GTX1+1)». Como hay valores que son 0, para poder aplicar la transformación es necesario sumar 1 a la variable.



Paso 12. Se realizan de nuevo los tests de Normalidad y homogenidad de varianzas siguiendo los pasos 4 a 6, pero usando la variable transformada «InvGTX1» en vez de «GTX1». Ahora se observa que en las dos especies, tanto en el test de Kolmogorov-Smirnov con la corrección de Lilliefors como en el test Shapiro-Wilk, la p > 0,05 y, por tanto, se acepta la hipótesis nula de que los datos se ajustan a una distribución Normal en ambas especies.

Pruebas de normalidad

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Especie	Estadístico	gl	Sig.	Estadístico	gl	Sig.
InvGTX1	A. minutum	,086	67	,200*	,964	67	,051
	A. tamarense	,085	67	,200*	,975	67	,189

a. Corrección de la significación de Lilliefors

*. Este es un límite inferior de la significación verdadera.

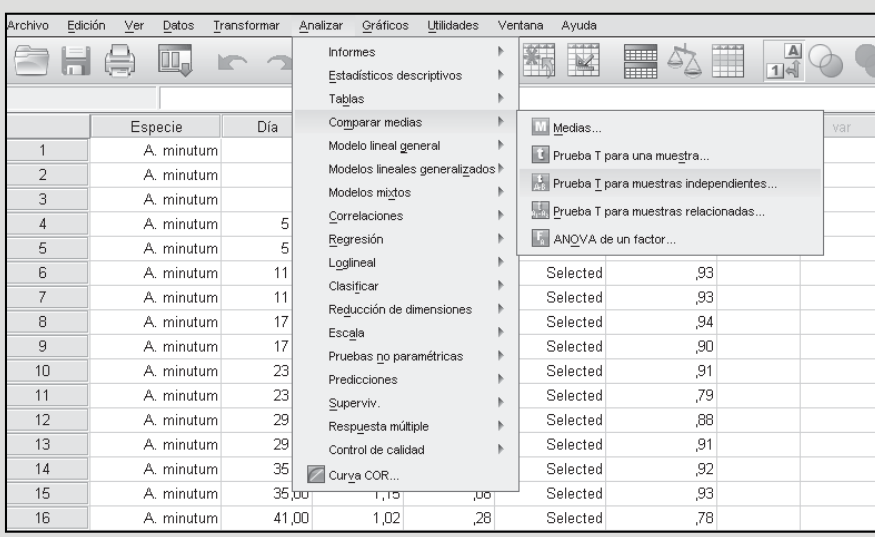
CUADRO IX.1. *(Continuación)*[SPSS]

Paso 13. Sin embargo, la prueba de Levene sigue mostrando que no hay homogeneidad de varianzas (p < 0,001). A pesar de que no se cumple este requisito, se puede usar el *t*-test porque es posible contemplar el supuesto de no homogeneidad.

Prueba de homogeneidad de la varianza

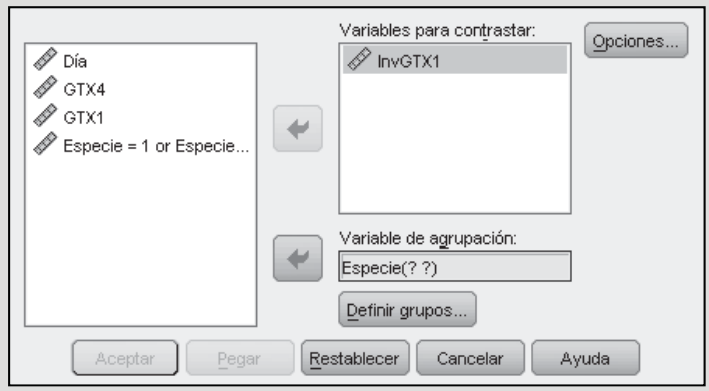
		Estadístico de Levene	gl1	gl2	Sig.
InvGTX1	Basándose en la media	15,679	1	132	,000
	Basándose en la mediana.	15,754	1	132	,000
	Basándose en la mediana y con gl corregido	15,754	1	113,555	,000
	Basándose en la media recortada	16,099	1	132	,000

Paso 14. Una vez comprobados nuestros supuestos de distribución Normal y homogeneidad podemos realizar el *t*-test, para lo cual regresamos al menú principal a «Analizar» ⇨ «Comparar medias» ⇨ «Prueba T para muestras independientes…».

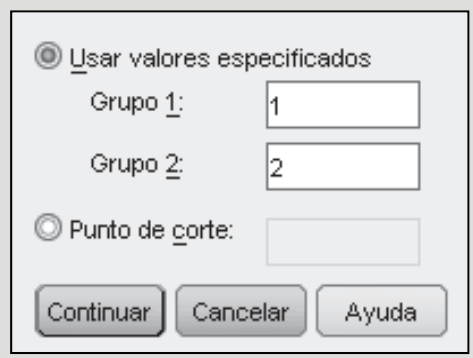


Paso 15. En la ventana que aparece, pasamos nuestra variable «InvGTX1» al cuadro «Variables para contrastar» y «Especie» a la sección de «Variable de agrupación». En este último paso debemos entrar en «Definir grupos», para especificar en la ventana correspondiente los grupos que queremos contrastar.

CUADRO IX.1. *(Continuación)*[SPSS]



Paso 16. En «Grupo 1:» se pone «1» que es la especie *A. minutum* y en «Grupo 2:» se pone «2» que es *A. tamarense*.



Paso 17. Como habíamos mencionado anteriormente, SPSS realiza el cálculo del *t*-test asumiendo o no la igualdad de las varianzas. Una vez analizados los resultados, como $p < 0{,}05$ rechazamos la H_o de igualdad de medias ($t = 5{,}7$, $df = 108{,}1$, $p < 0{,}001$), concluyendo que hay diferencias significativas entre las especies *A. minutum* y *A. tamarense* en la cantidad de toxina GTX1 por célula.

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias				
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia
InvGTX1	Se han asumido varianzas iguales	15,679	,000	5,730	132	,000	,06154	,01074
	No se han asumido varianzas iguales			5,730	108,144	,000	,06154	,01074

IX.1.3.2. Muestras dependientes

El *t*-test para muestras relacionadas compara las medias de dos variables de un solo grupo. Calcula las diferencias entre los valores de cada variable y contrasta si la diferencia media es significativamente distinta de cero (Sánchez, 1999).

Este test asume que las muestras son dependientes, pareadas o relacionadas, por tanto, un requisito fundamental es tener un número igual de observaciones en ambas variables.

Este *t*-test para muestras dependientes no exige ningún supuesto sobre las varianzas, pero sí requiere que la distribución de las diferencias de los valores de cada par sea Normal. Si no lo es, pero el tamaño de la muestra es grande, el teorema central del límite garantiza que la distribución de probabilidad de la diferencia de las medias sea aproximadamente Normal, lo que permite utilizar la distribución Normal en lugar de la distribución *t*. Dado que ambas distribuciones son prácticamente iguales cuando el número de grados de libertad es muy grande, esto significa que con muestras grandes podemos aplicar la prueba aunque la distribución de la variable se desvíe de la normalidad.

Las muestras dependientes aparecen a menudo cuando se evalúa una misma variable más de una vez en cada sujeto de la muestra (por ejemplo en intervalos de tiempo diferentes). El *t*-test no se centra en la variabilidad que puede darse entre los individuos, sino en las diferencias que se observan en un mismo sujeto entre un momento y otro (ver ejemplo en Cuadro IX.2). También se aplica este test en estudios de casos y controles donde cada caso se aparea individualmente con un control.

CUADRO IX.2. *t*-test para muestras dependientes

Aplicación con R.

EJEMPLO. En el Cuadro IX.1 se mostraron datos de peso y altura de niños y niñas entre los 2 y 4 años en Italia y España. El objetivo ahora es determinar si existen diferencias significativas en la altura de las niñas entre los 2 y 3 años en Italia. Los niños y las niñas que se van midiendo y tomando su peso conforme van cumpliendo años son siempre los mismos.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro IX.2.R**, que se muestra en la siguiente ventana. Sólo es necesario especificar lo que se solicita en la sección modificable por el usuario. Para que este *script* funcione correctamente es necesario tener instalados los paquetes "mvtnorm", "VGAM", "splines", "stats4", "lawstat" y "e1071", o realizar la instalación como se explica en el Apéndice II.

CUADRO IX.2. *(Continuación)*[R]

```
#Cuadro IX.2
#Test t para muestras dependientes.

################################################
#Sección modificable por el usuario
################################################

#Lectura de datos.
datos<-read.csv2("Cuadro IX.1.V.csv",header=TRUE,encoding="latin1")

#Selección variable dependiente. En caso de que se deseen utilizar val
#varInteres<-c("Peso")
varInteres<-c("Altura")

#Selección de un tipo de transformación para las variables de interés.
varInteres1X2<-NULL
varInteres1X<-NULL
varInteresLN<-NULL
varInteresLOG10<-NULL
varInteresSQR<-NULL
varInteresX2<-NULL
varInteresX3<-NULL
varInteresEX<-NULL
varInteresASN<-NULL

#varInteres1X2<-c(1,1)
#varInteres1X<-c(1,1)
#varInteresLN<-c(1,1)
#varInteresLOG10<-c(1,1)
#varInteresSQR<-c(1,1)
#varInteresX2<-c(1,1)
#varInteresX3<-c(1,1)
#varInteresEX<-c(1,1)
#varInteresASN<-c(1,1)
```

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro IX.1.V.csv**.

datos<-read.csv2("Cuadro IX.1.V.csv",header=TRUE,encoding="latin1")

Paso 2. Se selecciona la variable dependiente, en este caso la altura.

#varInteres<-c("Peso")
varInteres<-c("Altura")

Paso 3. En caso de que sea necesario es posible transformar la variable dependiente, para ello simplemente hay que quitar el símbolo # en el tipo de transformación que se quiera realizar. En primer lugar se hace el análisis sin transformar los datos, y en el caso de no cumplirse la Normalidad y/o homogeneidad de varianzas, entonces se realizaría el análisis aplicando la transformación más adecuada.

CUADRO IX.2. *(Continuación)*[R]

```
#varInteres1X2<-c(1,1)
#varInteres1X<-c(1,1)
#varInteresLN<-c(1,1)
#varInteresLOG10<-c(1,1)
#varInteresSQR<-c(1,1)
#varInteresX2<-c(1,1)
#varInteresX3<-c(1,1)
#varInteresEX<-c(1,1)
#varInteresASN<-c(1,1)
```

Paso 4. A continuación se indica cual es el factor en «varPoblacion» y se seleccionan los dos grupos que se van a comparar. En el caso de que sólo existan dos grupos, es posible poner la opción «NULL» sin necesidad de especificar los dos grupos a comparar. En el ejemplo, como se quiere determinar las diferencias entre los 2 y 3 años, se ponen esos dos grupos.

```
varPoblacion<-c("Edad")
varComparar1<-c("2")
varComparar2<-c("3")
#varComparar1<-NULL
```

Paso 5. También es posible seleccionar grupos en dos variables diferentes (un máximo de dos grupos en cada variable). En el ejemplo, como se quiere comparar solo las niñas de Italia, se seleccionan estos grupos en las variables «País» y «Sexo».

```
varSeleccion1<-list(c("País","Italia"))
#varSeleccion1<-NULL
varSeleccion2<-list(c("Sexo","Niña"))
#varSeleccion2<-NULL
```

Paso 6. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
ArchivodeSalida<-"Salida Cuadro IX.2.txt"
```

Paso 7. El test de Shapiro-Wilk muestra que las dos poblaciones de niñas de 2 (p = 0,415) y 3 años (p = 0,29) tienen distribución Normal. Además se cumple el requisito de homogeneidad de varianzas, tanto considerando la media (p = 0,831) como la mediana (p = 0,817). Por lo tanto, no es necesario realizar de nuevo el análisis transformando la variable dependiente. En el caso de que no se cumpliera alguno de los dos requisitos, también sale información sobre la asimetría y la curtosis de cada una de las poblaciones, lo cual ayudaría a elegir mejor la transformación a realizar.

CUADRO IX.2. *(Continuación)*[R]

```
[[1]]
datos3[, "poblaciones"]: 2

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9266, p-value = 0.4154

------------------------------------------------------------
datos3[, "poblaciones"]: 3

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9113, p-value = 0.2901


[[2]]

        classical Levene's test based on the absolute deviations from the mean
        ( none not applied because the location is not set to median )

data:  datos3$valores
Test Statistic = 0.0468, p-value = 0.8312


[[3]]

        modified robust Brown-Forsythe Levene-type test based on the absolute
        deviations from the median

data:  datos3$valores
Test Statistic = 0.0551, p-value = 0.817


[[4]]
                                        2                       3
            "Sesgo" "-0.280232427604095" "0.0737002993850063"

[[5]]
                                       2                       3
         "Curtosis" "-1.26371666194686" "-1.69531691801242"

[[6]]

        Paired t-test

data:  datos3$valores by datos3$poblaciones
t = -25.1073, df = 9, p-value = 1.212e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.938817 -7.461183
sample estimates:
mean of the differences
                  -8.2
```

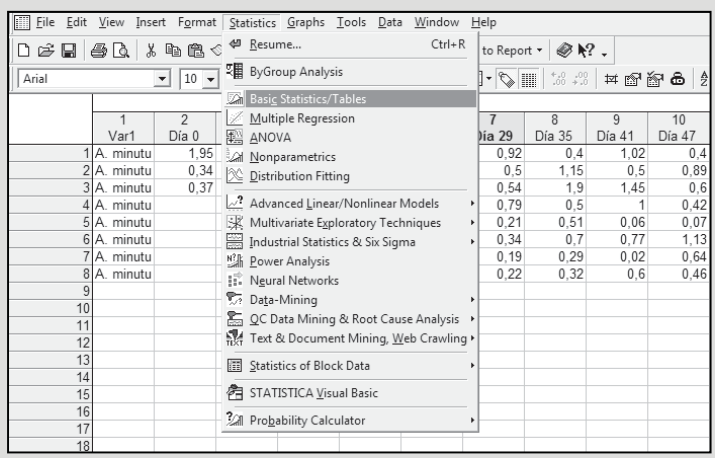
Por último, el test de la *t* muestra que existen diferencias significativas en la altura de las niñas italianas entre los 2 y 3 años (t =-25,1, df = 9, p < 0,001), con una diferencia de 8,2 cm.

CUADRO IX.2. *(Continuación)*^{STATISTICA}

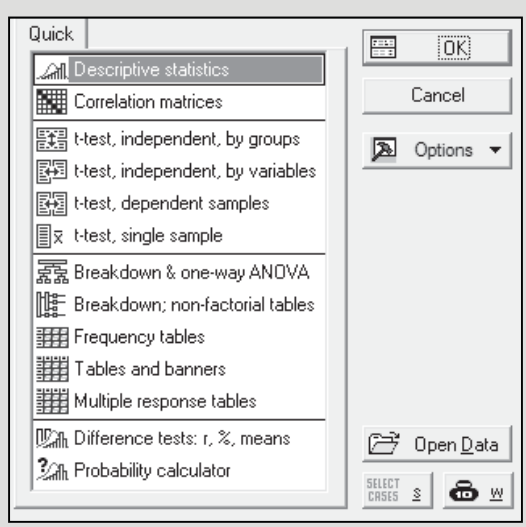
Aplicación con STATISTICA.

EJEMPLO. Vamos a seguir con el ejemplo del Cuadro IX.1, que era un estudio sobre la cantidad de dos toxinas, GTX4 y GTX1 (en pg célula^{-1}), en tres especies de fitoplancton tóxico del género *Alexandrium* a lo largo del tiempo: *A. minutum, A. catenella* y *A. tamarense*. Los datos se encuentran en el archivo **Cuadro IX.2.sta**. En este caso vamos a ver si existen diferencias significativas en la cantidad de la toxina GTX4 en la especie *A. minutum* entre los días 17 y 23 del experimento.

Paso 1. Los datos hay que organizarlos como se muestran en el archivo **Cuadro IX.2.sta**. El *t*-test para muestras pareadas no tiene como requisito la homogeneidad de varianzas, pero si el número de datos es pequeño, como en este caso, es necesario testar la Normalidad. Para ello, se selecciona «Statistics» ⇨ «Basic Statistics/Tables».

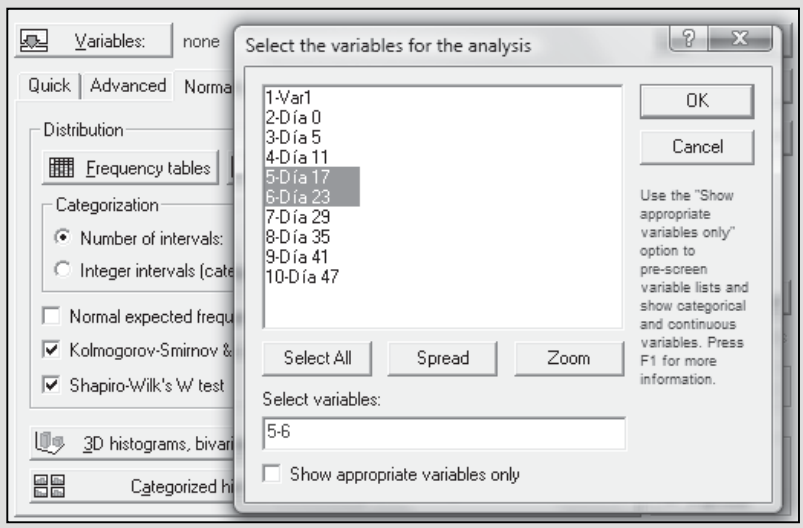

File Edit View Insert Format Statistics Graphs Tools Data Window Help
Resume... Ctrl+R
ByGroup Analysis
Basic Statistics/Tables
Multiple Regression
ANOVA
Nonparametrics
Distribution Fitting
Advanced Linear/Nonlinear Models
Multivariate Exploratory Techniques
Industrial Statistics & Six Sigma
Power Analysis
Neural Networks
Data-Mining
QC Data Mining & Root Cause Analysis
Text & Document Mining, Web Crawling
Statistics of Block Data
STATISTICA Visual Basic
Probability Calculator
Arial
10
1 Var1
2 Día 0
7 Día 29
8 Día 35
9 Día 41
10 Día 47
1 A. minutu 1,95 0,92 0,4 1,02 0,4
2 A. minutu 0,34 0,5 1,15 0,5 0,89
3 A. minutu 0,37 0,54 1,9 1,45 0,6
4 A. minutu 0,79 0,5 1 0,42
5 A. minutu 0,21 0,51 0,06 0,07
6 A. minutu 0,34 0,7 0,77 1,13
7 A. minutu 0,19 0,29 0,02 0,64
8 A. minutu 0,22 0,32 0,6 0,46

Paso 2. A continuación «Descriptive statistics».

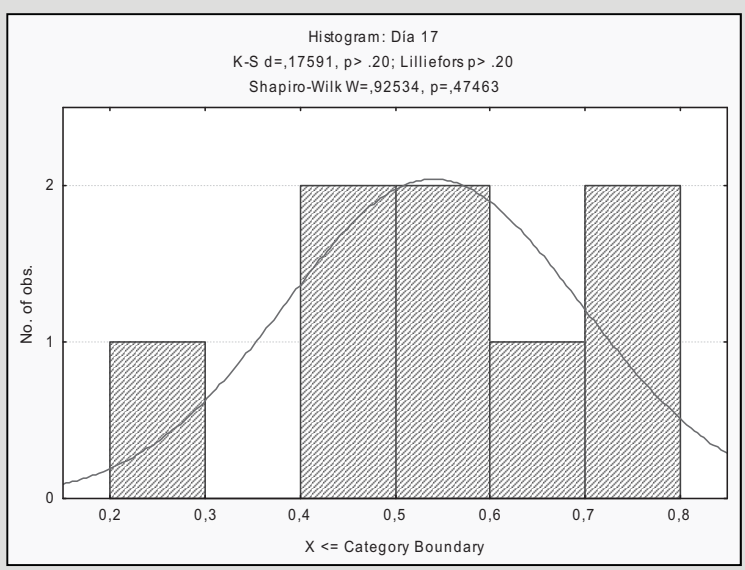


CUADRO IX.2. *(Continuación)*[STATISTICA]

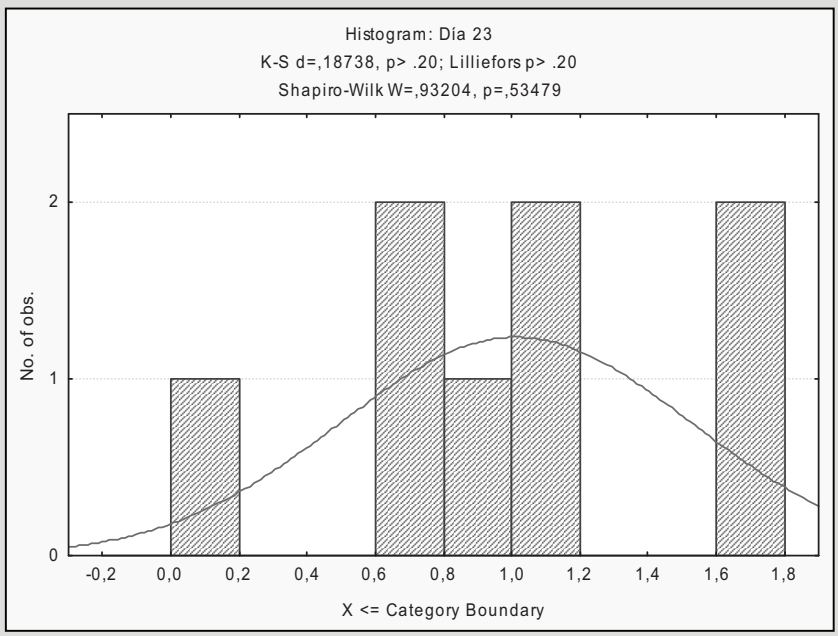
Paso 3. En la pestaña «Normality» se seleccionan los test «Kolmogorov-Smirnov & Lilliefors test for normality» y «Shapiro-Wilk´s W test». En «Variables» seleccionamos los días 17 y 23. Por último, pulsamos el botón «Histograms».



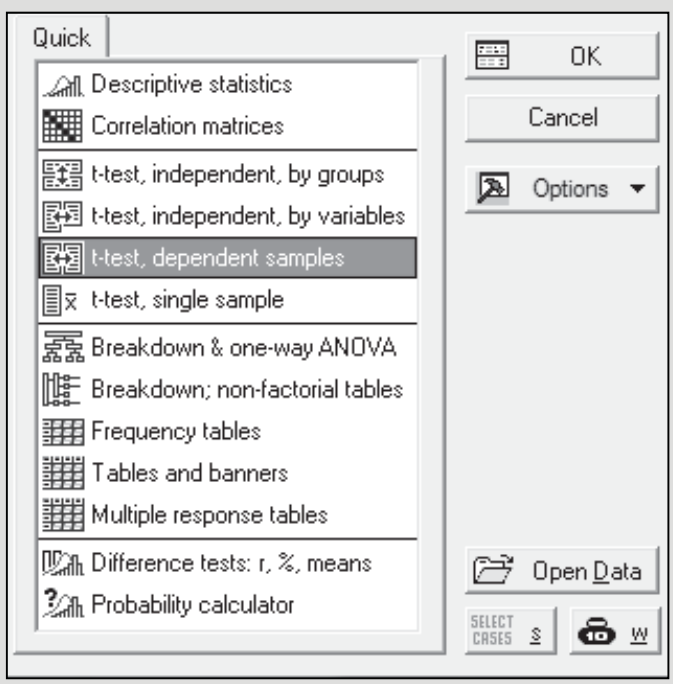
Paso 4. En ambos días las probabilidades del test Kolmogorov–Smirnov con la corrección de Lilliefors y Shapiro-Wilk son mayores que 0,05, por tanto, se acepta la hipótesis de bondad de ajuste y los datos se ajustan a una distribución Normal. En el caso de que la distribución no fuera Normal en alguno de los dos días del experimento sería necesario realizar una transformación como se explica en los pasos 10 y 11 del Cuadro IX.1 en la aplicación con STATISTICA.



CUADRO IX.2. *(Continuación)* *STATISTICA*

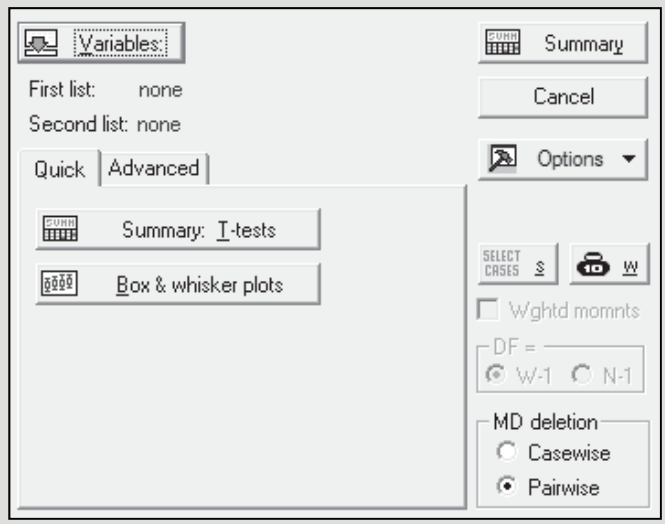


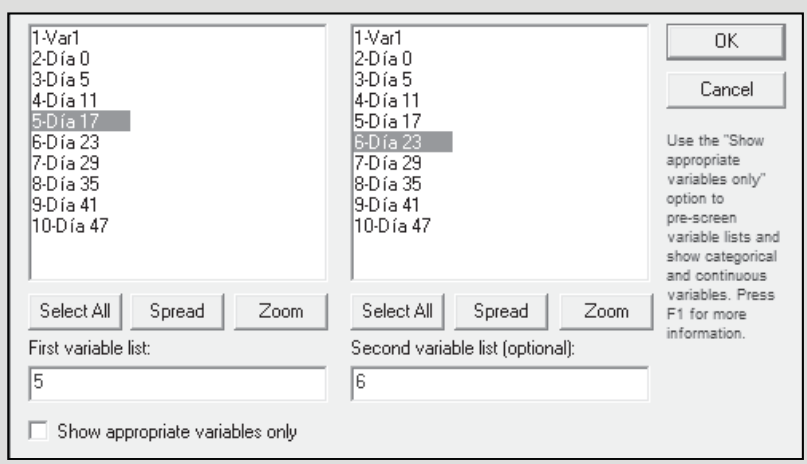
Paso 5. En la ventana del paso 2 se selecciona «t-test, dependent samples».



CUADRO IX.2. *(Continuación)*[STATISTICA]

Paso 6. En la siguiente ventana se pulsa en «Variables» y aparece otra ventana donde se seleccionan los días 17 y 23 del experimento sin importar el orden.





Paso 7. Al pulsar «Summary» aparecen los resultados donde se observa que no existen diferencias significativas en la concentración de toxina GTX4 en la especie *A. minutum* entre los días 17 y 23 del experimento (*t*-test = -2,17, *df* = 7, *p* = 0,067).

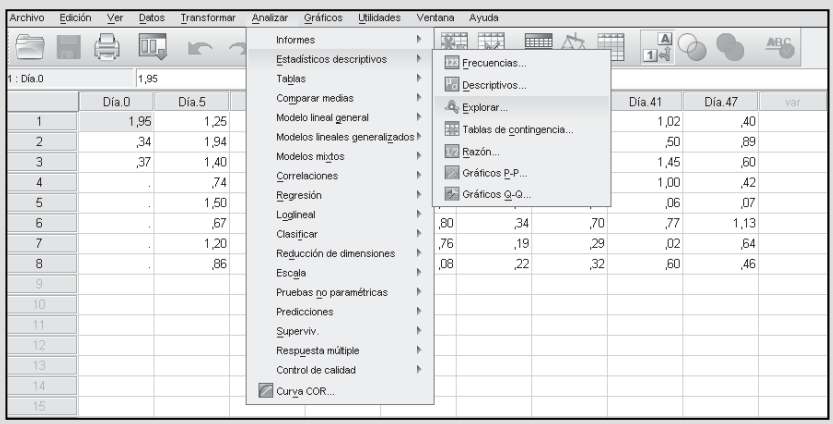
	T-test for Dependent Samples (Cuadro VIII.2.sta) Marked differences are significant at p < ,05000							
Variable	Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
Día 17	0,540000	0,156296						
Día 23	1,012500	0,516243	8	-0,472500	0,616598	-2,16743	7	0,066870

CUADRO IX.2. *(Continuación)*[SPSS]

Aplicación con SPSS.

EJEMPLO. Vamos a utilizar los mismos datos que en el caso de la estimación con STATISTICA, los cuales se encuentran en el archivo **Cuadro IX.2.sav**. En este caso vamos a ver si existen diferencias significativas en la cantidad de la toxina GTX4 en la especie *A. minutum* entre los días 5 y 17, y entre los días 5 y 47 del experimento.

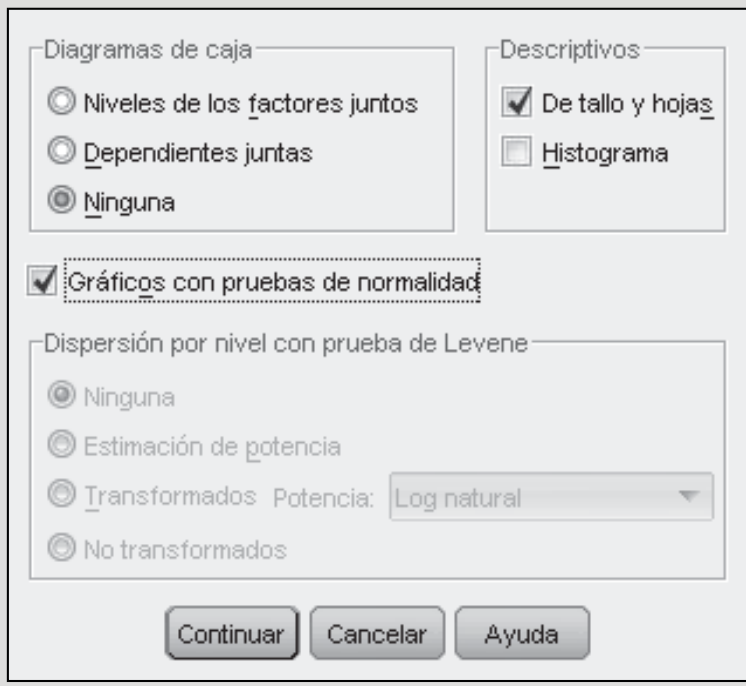
Paso 1. Los datos hay que organizarlos como se muestran en el archivo **Cuadro IX.2.sav**. El *t*-test para muestras pareadas no tiene como requisito la homogeneidad de varianzas, pero si el número de datos es pequeño, como en este caso, es necesario testar la Normalidad. Se selecciona «Analizar» ⇨ «Estadísticos descriptivos…» ⇨ «Explorar…».



Paso 2. En la ventana que aparece, en «Lista de dependientes» se incluyen las variables a las que se va a determinar si presentan una distribución Normal, en nuestro ejemplo «Día.5», «Día.17» y «Día.47».En «Visualización» se selecciona «Ambos» y, por último se pulsa el botón «Gráficos».

CUADRO IX.2. *(Continuación)*[SPSS]

Paso 3. A continuación en «Diagrama de cajas» se selecciona «Ninguna» y se marca «Gráficos con pruebas de normalidad».



Paso 4. En los tres días las probabilidades del test Kolmogorov–Smirnov con la corrección de Lilliefors y Shapiro-Wilk son mayores que 0,05, por tanto, se acepta la hipótesis de bondad de ajuste y los datos se ajustan a una distribución Normal. En el caso de que la distribución no fuera Normal en alguno de los días del experimento sería necesario realizar una transformación como se explica en los pasos 10 y 11 del Cuadro IX.1 en la aplicación con SPSS.

Pruebas de normalidad

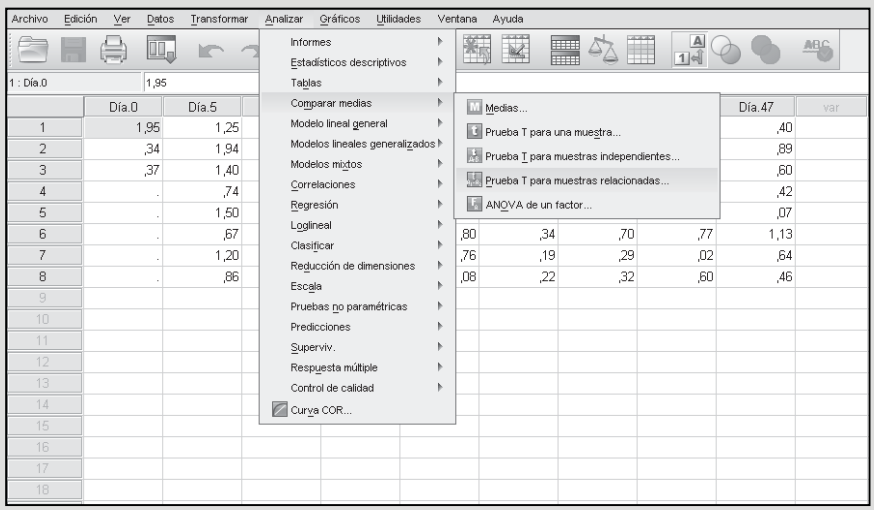
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Día.5	,158	8	,200*	,950	8	,712
Día.17	,176	8	,200*	,925	8	,475
Día.47	,172	8	,200*	,966	8	,868

a. Corrección de la significación de Lilliefors

*. Este es un límite inferior de la significación verdadera.

CUADRO IX.2. *(Continuación)*[SPSS]

Paso 5. Se selecciona «Analizar» ⇨ «Comparar medias» ⇨ «Prueba T para muestras relacionadas…».



Paso 6. Se especifican las variables a comparar; en este caso el día 5 con el día 17 y el día 5 con el día 47 del experimento.

Paso 7. Los resultados muestran que existen diferencias significativas en la concentración de toxina GTX4 en la especie *A. minutum* entre los días 5 y 17 del experimento (t-test = 3,896, df = 7, p = 0,006) y entre los días 5 y 47 del experimento (t-test = 3,097, df = 7, p = 0,017).

Prueba de muestras relacionadas

		t	gl	Sig. (bilateral)
Par 1	Día.5 - Día.17	3,896	7	,006
Par 2	Día.5 - Día.47	3,097	7	,017

IX.1.4. ANÁLISIS DE VARIANZA

En general, al realizar una investigación nos preguntamos si nuestras muestras, que consideramos independientes, pertenecen a una misma población. Los valores de las muestras comúnmente difieren, el problema sería determinar si a pesar de estas diferencias las poblaciones son iguales, y las variaciones son debidas al azar como consecuencia de la aleatoriedad de nuestro muestreo.

La prueba paramétrica más usada para verificar si dos o más medias muestrales proceden de la misma población es el análisis de varianza (ANOVA). El ANOVA al ser un método general, puede ser extendido a más de dos muestras y se puede demostrar que coincide con la *t* de Student si sólo se tienen dos muestras (Salvarrey, 2000; Azzimonti, 2003).

Para la aplicación de un ANOVA es necesario cumplir los siguientes supuestos:

- Las muestras son aleatorias.
- Las muestras son independientes.
- La distribución de la población de donde fueron extraídas es Normal.
- Las varianzas son iguales en cada grupo o nivel de los factores considerados.

El ANOVA calcula la variación de las medias y estima la variación "natural" de la población, para luego hacer una comparación entre ambas. La variación natural se mide por medio de la "intravarianza" o "variación del error". Si las muestras son de una misma población, la varianza de las medias es la enésima parte de la varianza de la población. Si se descarta esa situación es porque, aparte del azar, algo más diferencia a las muestras. Si las diferencias son sólo debidas al azar, las dos varianzas corregidas por sus grados de libertad son del mismo orden y su cociente vale más o menos 1. El cociente de varianzas corregidas tiene distribución *F*, por lo que se busca en las tablas de la distribución *F* si las relaciones obtenidas son aceptables como cercanas a 1 o no (Salvarrey, 2000).

El análisis de varianza se puede realizar con una variable dependiente (ANOVA univariante) o con varias variables dependientes (ANOVA multivariante o MANOVA). También es posible realizar el análisis de varianza de un factor (ANOVA unifactorial) o con varios factores (ANOVA multifactorial).

El llamado modelo lineal general (MGL) univariante realiza un análisis de regresión y uno de varianza para una variable dependiente mediante uno o más factores o variables independientes (los factores dividen la población en grupos). Con este procedimiento podemos contrastar hipótesis acerca de los efectos de una o más variables (factores o tratamientos) sobre la media de una única variable dependiente. Este análisis exige que la variable dependiente sea cuantitativa y los factores categóricos, es decir, que pueden tener valores numéricos o de cadena (letras).

En el Cuadro IX.3 se explica como se realiza un ANOVA unifactorial y multifactorial.

CUADRO IX.3. Análisis de varianza unifactorial y multifactorial

Aplicación con R.

EJEMPLO. En el archivo **Cuadro IX.3.V.csv** hay datos biométricos de dos especies de tiburones (*Scyliorhinus canicula* y *Galeus melastomus*), los cuales fueron tomados por dos investigadores en cuatro zonas (dos en el Mediterráneo y dos en el Atlántico). El objetivo es determinar si hay diferencias significativas en la longitud de la aleta anal entre las especies, zonas e investigadores.

Después de haber invocado al programa R y seleccionado el directorio, cómo se explica en el Apéndice II, abrimos el archivo **Cuadro IX.3.R**, que se muestra en la siguiente ventana. Sólo es necesario especificar lo que se solicita en la sección modificable por el usuario. Para que este *script* funcione correctamente es necesario tener instalados los paquetes "nortest", "mvtnorm", "class", "VGAM", "splines", "stats4", "lawstat", "car" "multcomp", "survival" y "e1071", o realizar la instalación como se explica en el Apéndice II.

```
#Cuadro IX.3
#Análisis de varianza unifactorial y multifactorial.

################################################
#Sección modificable por el usuario
################################################

#Lectura de datos.
datos<-read.csv2("Cuadro IX.3.V.csv",header=TRUE,encoding="latin1")

#Selección variable dependiente. En caso de que se deseen utilizar val
varInteres<-c("Longitud.anal")
#varInteres<-c("Diámetro.ojo")

#Selección de un tipo de transformación para las variables de interés.
varInteres1X2<-NULL
varInteres1X<-NULL
varInteresLN<-NULL
varInteresLOG10<-NULL
varInteresSQR<-NULL
varInteresX2<-NULL
varInteresX3<-NULL
varInteresEX<-NULL
varInteresASN<-NULL

#varInteres1X2<-c(1,1)
#varInteres1X<-c(1,1)
#varInteresLN<-c(1,1)
#varInteresLOG10<-c(1,1)
#varInteresSQR<-c(1,1)
#varInteresX2<-c(1,1)
#varInteresX3<-c(1,1)
#varInteresEX<-c(1,1)
#varInteresASN<-c(1,1)
```

Para abordar el objetivo planteado es necesario hacer un ANOVA multifactorial, es decir, considerar los tres factores («Especie», «Investigador» y «Zona»). Sin embargo, para explicar cómo se hace un ANOVA con un factor (unifactorial), en primer lugar vamos a trabajar solo con el factor «Especie» y luego realizaremos el análisis considerando los tres factores. Si se quiere hacer directamente un ANOVA multifactorial, éste se describe a partir del paso 11.

CUADRO IX.3. *(Continuación)*[R]

ANOVA unifactorial

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro IX.3.V.csv**.

```
datos<-read.csv2("Cuadro IX.3.V.csv",header=TRUE,encoding="latin1")
```

Paso 2. Se selecciona la variable dependiente, en este caso la longitud de la aleta anal.

```
varInteres<-c("Longitud.anal")
#varInteres<-c("Diámetro.ojo")
```

Paso 3. En caso de que sea necesario es posible transformar la variable dependiente, para ello simplemente hay que quitar el símbolo # en el tipo de transformación que se quiera realizar. En primer lugar se hace el análisis sin transformar los datos, y en el caso de no cumplirse la Normalidad y/o homogeneidad de varianzas, entonces se realizaría el análisis aplicando la transformación más adecuada.

```
#varInteres1X2<-c(1,1)
#varInteres1X<-c(1,1)
#varInteresLN<-c(1,1)
#varInteresLOG10<-c(1,1)
#varInteresSQR<-c(1,1)
#varInteresX2<-c(1,1)
#varInteresX3<-c(1,1)
#varInteresEX<-c(1,1)
#varInteresASN<-c(1,1)
```

Paso 4. A continuación se indica cuales son los factores con los que se va a trabajar. Es posible hacer el análisis con un máximo de 4 factores con este *script*. En este primer caso, como solo trabajamos con el factor «Especie», los otros factores tienen activada la opción «NULL» ya que no tienen el símbolo # delante de la instrucción.

```
Factor1<-c("Especie")
Factor2<-c("Investigador")
Factor2<-NULL
Factor3<-c("Zona")
Factor3<-NULL
Factor4<-c("")
Factor4<-NULL
```

Paso 5. También es posible seleccionar grupos en dos variables diferentes (un máximo de dos grupos en cada variable). En este ejemplo no se hace ninguna selección.

```
varSeleccion1<-list(c("Investigador","1"))
varSeleccion1<-NULL
varSeleccion2<-list(c("Zona","Atlántico-1","Atlántico-2"))
varSeleccion2<-NULL
```

CUADRO IX.3. *(Continuación)*[R]

Paso 6. Cuando existen varios factores, la descomposición de la suma de cuadrados en partes atribuidas a cada uno de los factores no es única. Se suelen utilizar cuatro tipos de sumas de cuadrados: tipo I para modelos anidados, tipo II para modelos equilibrados, tipo III para modelos no equilibrados (o también equilibrados), y tipo IV para modelos con casillas vacías. El tipo III es adecuado para la mayoría de las aplicaciones, por lo que será el que utilizaremos por defecto. Esta descripción es muy simplificada; los lectores interesados deberán consultar bibliografía complementaria para un uso correcto del ANOVA con modelos complejos.

```
#SS<-c(2)
SS<-c(3)
```

Paso 7. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
ArchivodeSalida<-"Salida Cuadro IX.3.txt"
```

Paso 8. En primer lugar nos fijamos en la Normalidad de los residuos y la homogeneidad de varianzas. Se observa que tanto el test Kolmogorov–Smirnov con la corrección de Lilliefors (p = 0,062), como el de Shapiro-Wilk (p = 0,062) tienen un valor p > 0,05 y, por tanto, los residuos se ajustan a una distribución Normal. El test de Levene considerando la mediana muestra que también se cumple la homogeneidad de varianzas (p = 0,098). Por lo tanto, no es necesario transformar los datos.

```
[[4]]

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.1131, p-value = 0.06213


[[5]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9614, p-value = 0.06199


[[6]]
[[6]][[1]]
[1] "Factor1"

[[6]][[2]]
Levene's Test for Homogeneity of Variance
      Df F value  Pr(>F)
group  1  2.8174 0.09882 .
      56
```

Paso 9. Si fuera necesario transformar los datos nos fijaríamos en la asimetría y la curtosis, y en base a la Figura IX.1 se escogería la mejor transformación. Se repetiría el análisis transformando los datos y, para ello, en el paso 3 quitaríamos el símbolo # de la transformación que queremos realizar.

CUADRO IX.3. *(Continuación)*[R]

Paso 10. Los resultados muestran que existen diferencias significativas en la longitud de la aleta anal entre las dos especies de tiburones (ANOVA, $F_{1, 56} = 27{,}5$, $p < 0{,}001$).

```
Response: datos2$valores
             Sum Sq Df F value    Pr(>F)
(Intercept) 26810.5  1 230.016 < 2.2e-16 ***
datos2$F1    3205.7  1  27.503 2.480e-06 ***
Residuals    6527.3 56
```

ANOVA multifactorial

Paso 11. Ahora se va a trabajar con los tres factores «Especie», «Investigador» y «Zona». La única modificación que hay que realizar en el *script* es incluir los dos nuevos factores y, para ello, ponemos el símbolo # antes de la instrucción «NULL» de cada factor.

```
Factor1<-c("Especie")
Factor2<-c("Investigador")
#Factor2<-NULL
Factor3<-c("Zona")
#Factor3<-NULL
Factor4<-c("")
Factor4<-NULL
```

Paso 12. Los residuos se ajustan a una distribución Normal: test Kolmogorov–Smirnov con la corrección de Lilliefors ($p = 0{,}345$) y el de Shapiro-Wilk ($p = 0{,}108$), y existe homogeneidad de varianzas para el factor 1 ($p = 0{,}251$), para el factor 2 ($p = 0{,}777$) y para el factor 3 ($p = 0{,}084$). Por lo tanto, no es necesario transformar los datos.

```
[[4]]

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.0865, p-value = 0.3453


[[5]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9664, p-value = 0.1078


[[6]]
[[6]][[1]]
[1] "Factor1"

[[6]][[2]]
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  1   1.344 0.2513
      56
```

CUADRO IX.3. *(Continuación)*[R]

Paso 13. A pesar de considerar el efecto «Investigador» y «Zona», siguen existiendo claras diferencias en la longitud de la aleta anal entre especies (ANOVA, $F_{1, 42}$ = 35,6, p < 0,001). Sin embargo, no existen diferencias significativas entre investigadores (ANOVA, $F_{1, 42}$ = 0,22, p = 0,638), es decir, no existen diferencias de criterio a la hora de tomar la medida. Tampoco existen diferencias entre zonas (ANOVA, $F_{3, 42}$ = 0,97, p = 0,417), En la tabla de ANOVA se muestra también el efecto conjunto de los factores, considerado como un factor adicional que recoge la posibilidad de que –además del efecto de la especie, investigador y la zona, que se suman– aparezca otro más, que también se suma a los anteriores, que es la interacción entre ellos. Se observa que algunas interacciones son significativas, como «Especie*Investigador» (ANOVA, $F_{1, 42}$ = 7,73, p = 0,008), pero no debe tenerse en cuenta una interacción cuando alguno de los factores que la forman no es significativo.

```
                           Sum Sq Df  F value    Pr(>F)
(Intercept)               27329.4  1 579.0386 < 2.2e-16 ***
datos2$F1                  1682.3  1  35.6438 4.383e-07 ***
datos2$F2                    10.6  1   0.2240  0.638466
datos2$F3                   137.0  3   0.9678  0.416922
datos2$F1:datos2$F2         364.9  1   7.7320  0.008086 **
datos2$F1:datos2$F3        3229.7  3  22.8098 6.407e-09 ***
datos2$F2:datos2$F3         432.1  3   3.0516  0.038815 *
datos2$F1:datos2$F2:datos2$F3 290.1  3   2.0487  0.121602
Residuals                  1982.3 42
```

Paso 14. Como en el caso del factor «Zona» hay 4 niveles, se podrían hacer pruebas *post hoc* para averiguar cuales son los niveles distintos. Sin embargo, no tiene sentido hacer comparaciones *post hoc* con un factor que no es significativo, ya que si una sola comparación fuera significativa también lo sería el factor. De todas formas R muestra los resultados post *hoc* del test Tukey, incluso para aquellos factores que solo tienen 2 niveles. Al fijarnos en el Factor 3, que es la «Zona», como era de esperar, se observa que no existen diferencias significativas al comparar cada una de las zonas: zona 2 con la 1 (p = 0,986), zona 3 con la 1 (p = 0,919), zona 4 con la 1 (p = 0,952), zona 3 con la 2 (p = 0,595), zona 4 con la 2 (p = 0,703), y zona 4 con la 3 (p = 1).

```
Linear Hypotheses:
                     Estimate Std. Error t value Pr(>|t|)
Factor1: 2 - 1 == 0 11.44521    1.91704   5.970   <1e-04 ***
Factor2: 2 - 1 == 0  0.90729    1.91704   0.473    0.995
Factor3: 2 - 1 == 0 -1.46708    2.49878  -0.587    0.986
Factor3: 3 - 1 == 0  2.32542    2.59529   0.896    0.919
Factor3: 4 - 1 == 0  2.24375    2.84818   0.788    0.952
Factor3: 3 - 2 == 0  3.79250    2.56672   1.478    0.595
Factor3: 4 - 2 == 0  3.71083    2.82217   1.315    0.703
Factor3: 4 - 3 == 0 -0.08167    2.90797  -0.028    1.000
```

CUADRO IX.3. *(Continuación)*[STATISTICA]

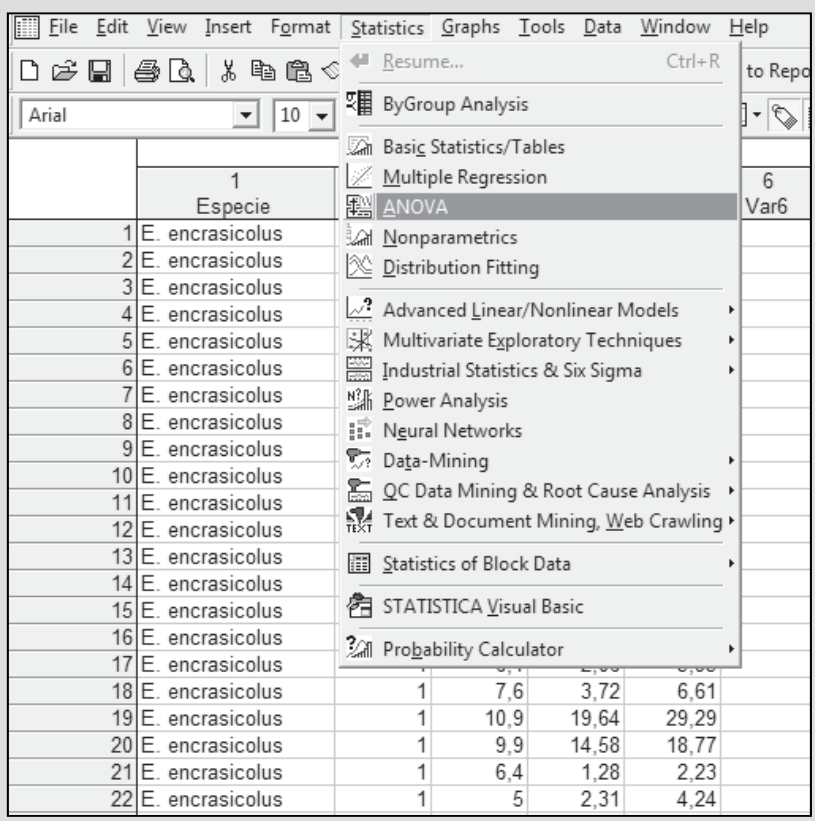
Aplicación con STATISTICA.

EJEMPLO. En el archivo **Cuadro IX.3.sta** hay datos del tamaño de las larvas de dos especies de peces de la familia Clupeidae, *Sardinella aurita* y *Engraulis encrasicolus*, y la cantidad de dos aminoácidos (Valina y Lisina) en pg larva^{-1}, en diferentes zonas del Canal de Sicilia. El objetivo es determinar si existen diferencias significativas en la cantidad del aminoácido Lisina entre las especies y/o las zonas.

Para abordar el objetivo planteado es necesario hacer un ANOVA multifactorial, es decir, considerar los dos factores («Especie» y «Zona»). Sin embargo, para explicar como se hace un ANOVA con un factor (unifactorial), en primer lugar vamos a trabajar solo con el factor «Especie» y luego realizaremos el análisis considerando los dos factores («Especie» y «Zona»). Si se quiere hacer directamente un ANOVA multifactorial, éste se describe a partir del paso 18.

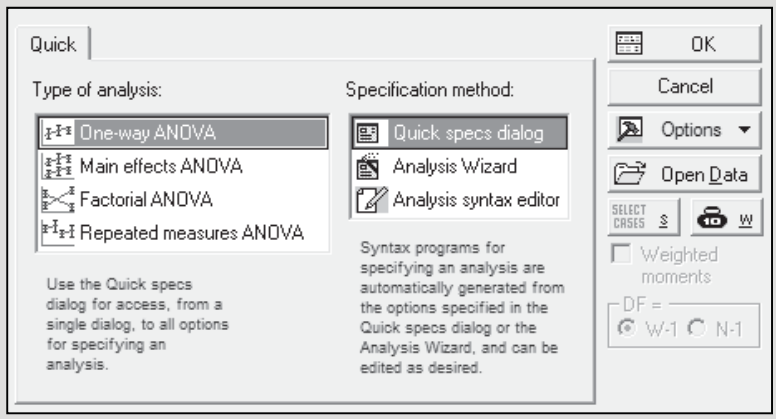
ANOVA unifactorial

Paso 1. Se selecciona «Statistics» ⇨ «ANOVA».

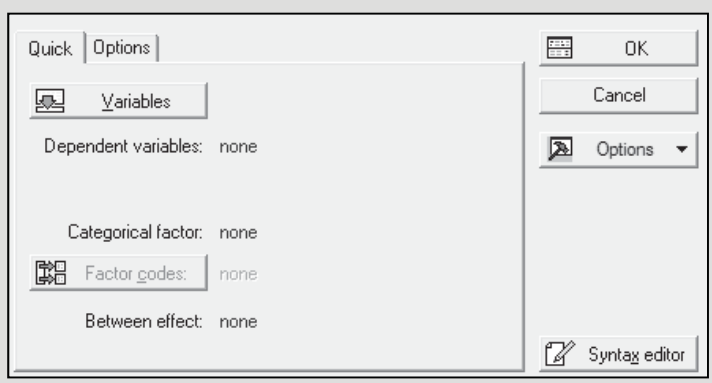


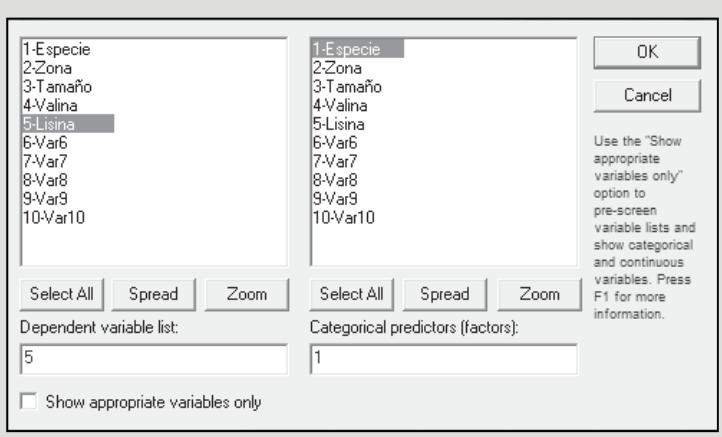
CUADRO IX.3. *(Continuación)*[STATISTICA]

Paso 2. En «Type of analysis» se selecciona «One-way ANOVA» y en «Specification method:» marcamos «Quick specs dialog».


Quick
Type of analysis:
One-way ANOVA
Main effects ANOVA
Factorial ANOVA
Repeated measures ANOVA
Use the Quick specs dialog for access, from a single dialog, to all options for specifying an analysis.
Specification method:
Quick specs dialog
Analysis Wizard
Analysis syntax editor
Syntax programs for specifying an analysis are automatically generated from the options specified in the Quick specs dialog or the Analysis Wizard, and can be edited as desired.
OK
Cancel
Options
Open Data
Weighted moments
DF =
W-1
N-1

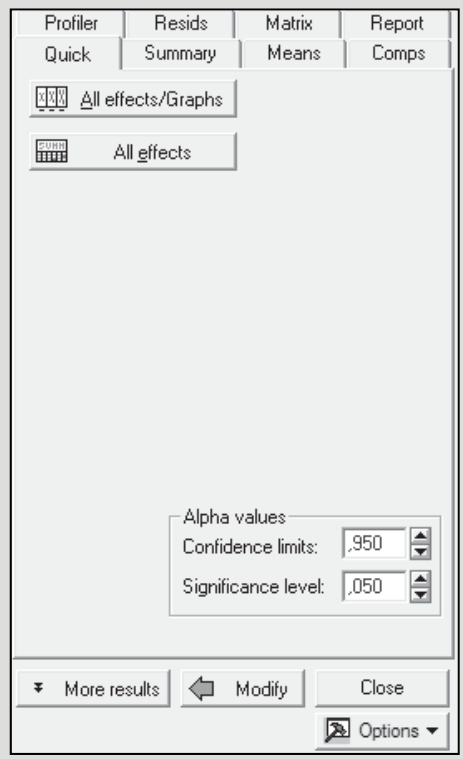
Paso 3. En la siguiente ventana, en la pestaña «Quick» se pulsa en «Variables» y aparece otra ventana donde, en «Dependent variable list» introducimos «Lisina», y en «Categorical predictor (factors):» introducimos el factor «Especie».

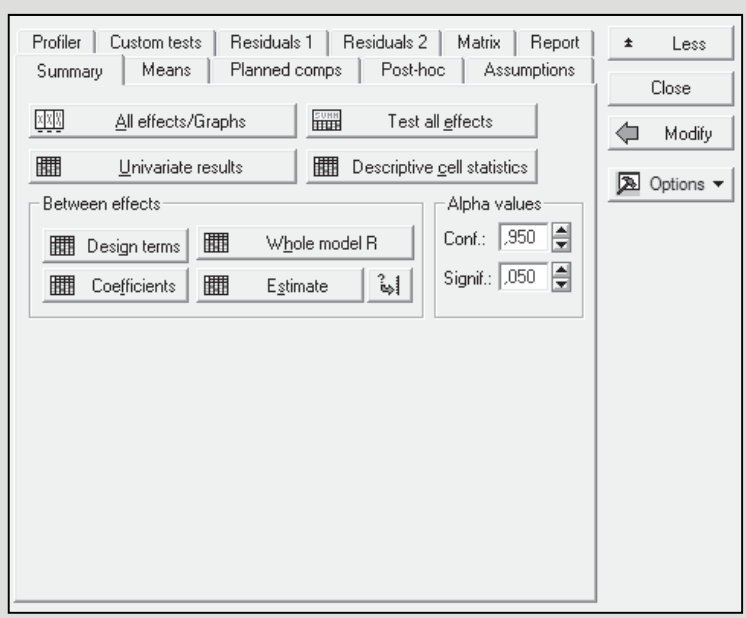




CUADRO IX.3. *(Continuación)*[STATISTICA]

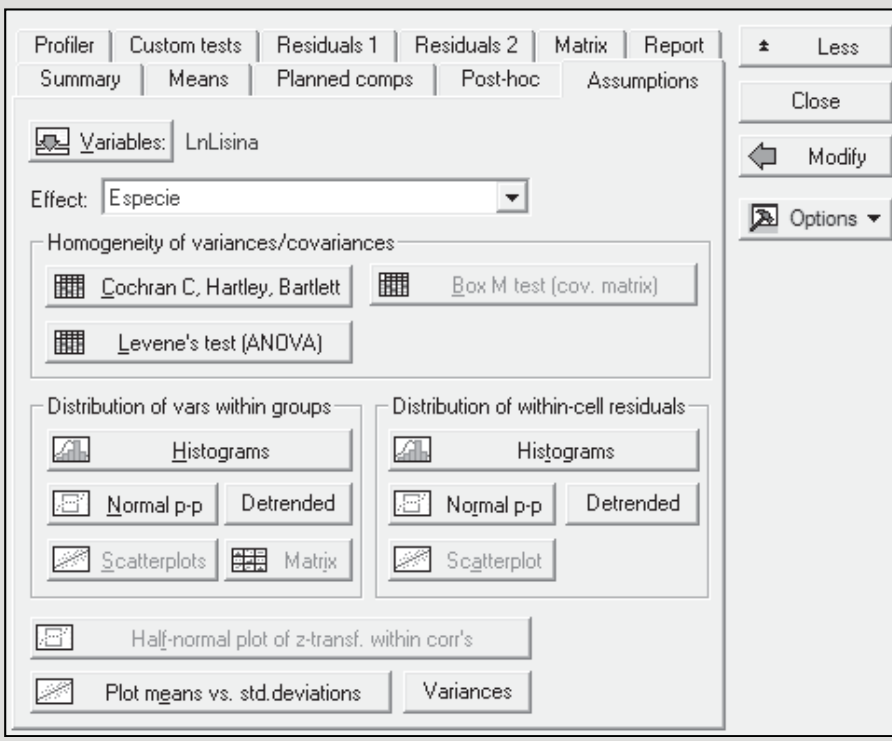
Paso 4. Al pulsar «OK» en el paso anterior, obtenemos la siguiente ventana donde pulsamos en «More results» para que aparezca la ventana extendida que se muestra a continuación.





CUADRO IX.3. *(Continuación)*[STATISTICA]

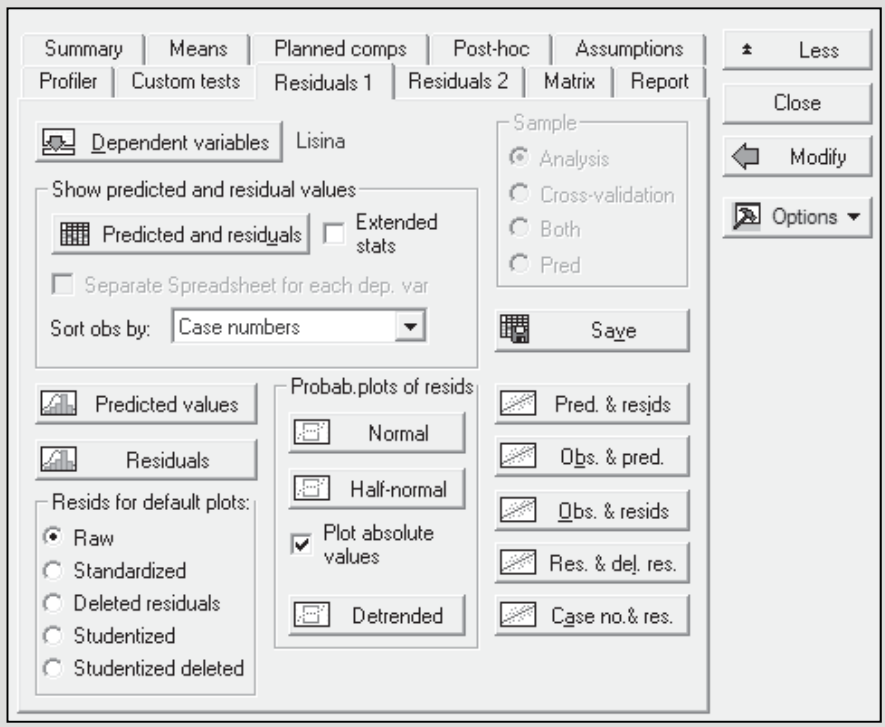
Paso 5. En la pestaña «Assumptions» pulsamos «Levene´s test (ANOVA)». Aparece la tabla que se muestra a continuación. Como la probabilidad es mayor que 0,05, aceptamos la hipótesis nula de homogeneidad de varianzas (p = 0,809). Por lo tanto, se cumple este requisito necesario para dar validez a los resultados que se obtienen con el ANOVA.



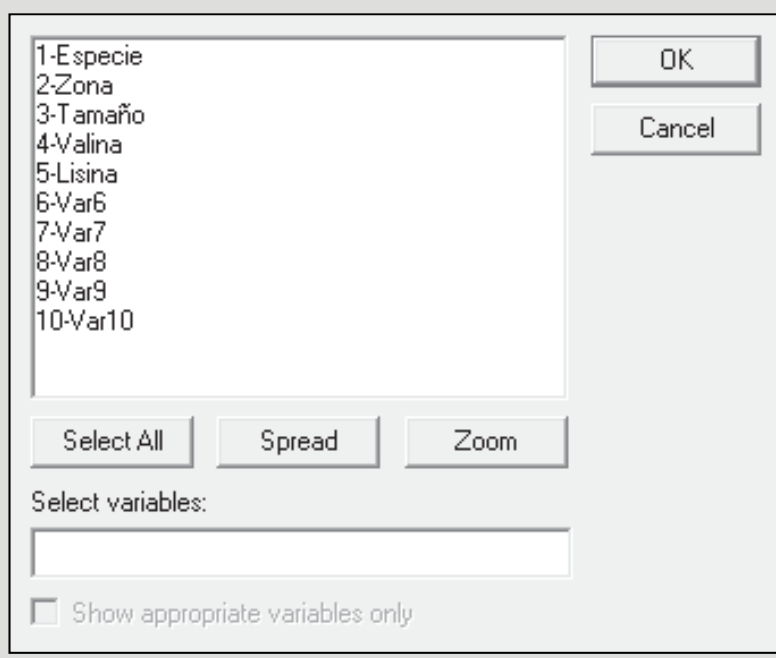
	Levene's Test for Homogeneity of Variances (Cuadro VIII.3.sta) Effect: Especie Degrees of freedom for all F's: 1, 236			
	MS Effect	MS Error	F	p
Lisina	2,736063	46,79568	0,058468	0,809143

Paso 6. Para comprobar la Normalidad, lo que se hace es ver si los residuos tienen distribución Normal. En la pestaña «Residuals 1» pulsamos «Save», como se muestra en la siguiente ventana.

CUADRO IX.3. *(Continuación)* STATISTICA



Paso 7. En la ventana que aparece pulsamos «OK» sin seleccionar ninguna variable.

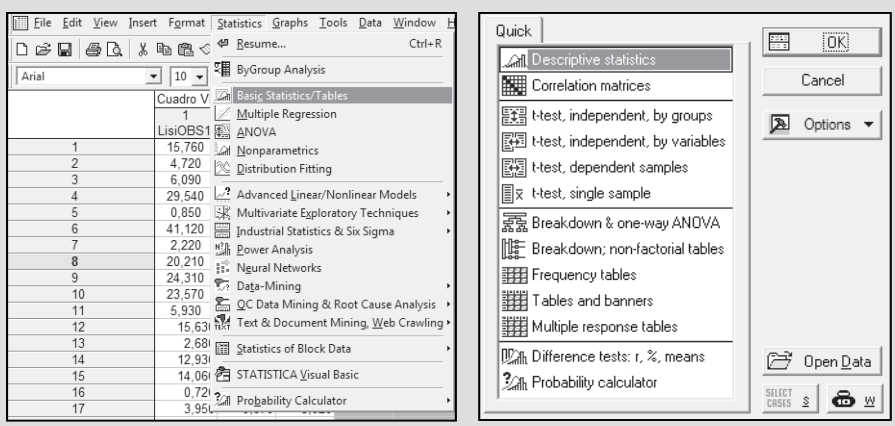


CUADRO IX.3. *(Continuación)* *STATISTICA*

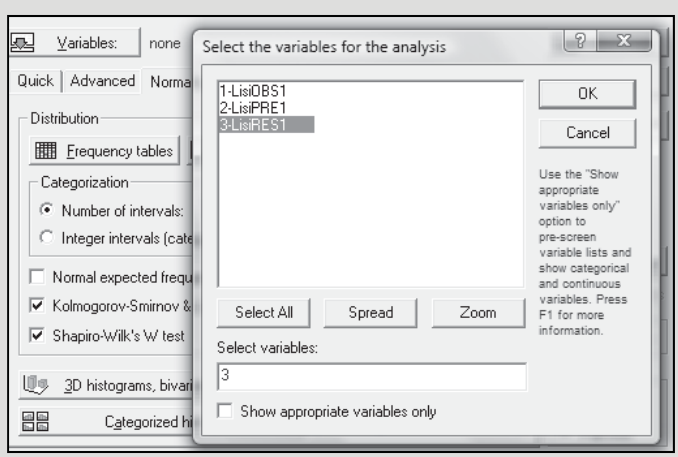
Paso 8. Se obtiene una tabla, en la cual la tercera columna son los residuos.

	Cuadro VIII.3.sta		
	1 LisiOBS1	2 LisiPRE1	3 LisiRES1
1	15,760	9,870	5,890
2	4,720	9,870	-5,150
3	6,090	9,870	-3,780
4	29,540	9,870	19,670
5	0,850	9,870	-9,020
6	41,120	9,870	31,250

Paso 9. Para determinar si los residuos tienen distribución Normal se selecciona «Statistics» ⇨ «Basic Statistics/Tables» ⇨ «Descriptive statistics».

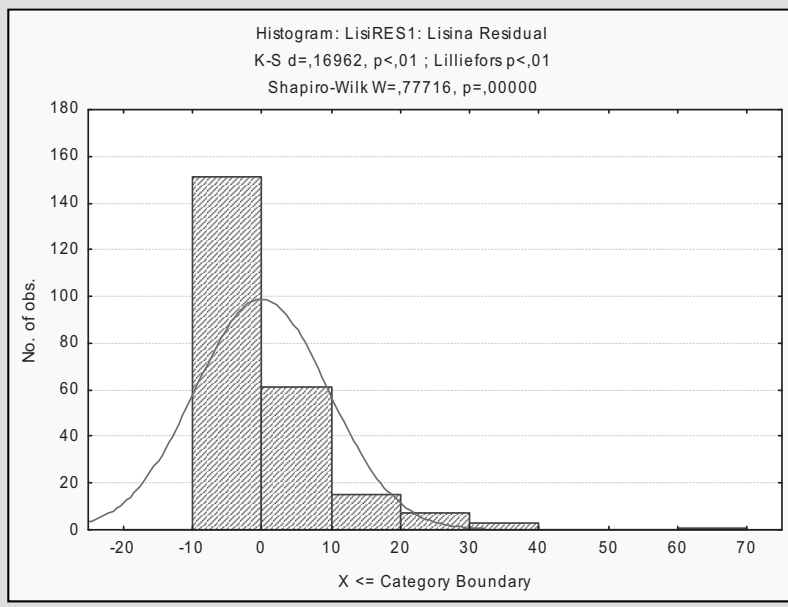


Paso 10. En la pestaña «Normality» marcamos los test «Kolmogorov-Smirnov & Lilliefors test for normality» y «Shapiro-Wilk´s test». En «Variables» se selecciona «LisiRES1».

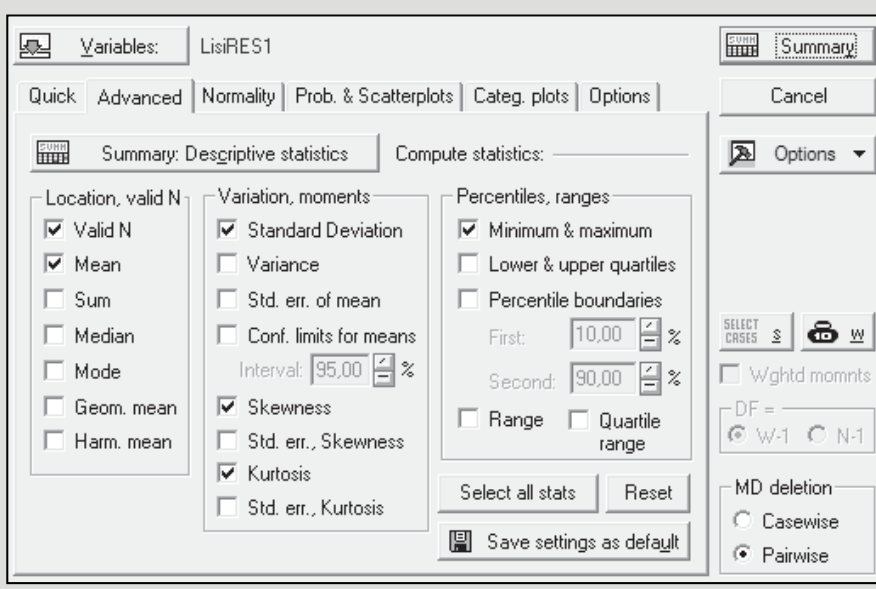


CUADRO IX.3. *(Continuación)*^{STATISTICA}

Paso 11. En la ventana anterior, al pulsar «Histograms» aparece el siguiente gráfico donde se observa que tanto el test Kolmogorov–Smirnov con la corrección de Lilliefors, como el de Shapiro-Wilk muestran una $p < 0{,}01$ y, por tanto, se rechaza la hipótesis nula, y los residuos no se ajustan a una distribución Normal. Por lo tanto, es necesario transformar los datos.


Histogram: LisiRES1: Lisina Residual
K-S d=,16962, p<,01 ; Lilliefors p<,01
Shapiro-Wilk W=,77716, p=,00000
No. of obs.
X <= Category Boundary

Paso 12. Para poder saber cual es la transformación más adecuada, en la pestaña «Advanced» marcamos «Skewness» y «Kurtosis».

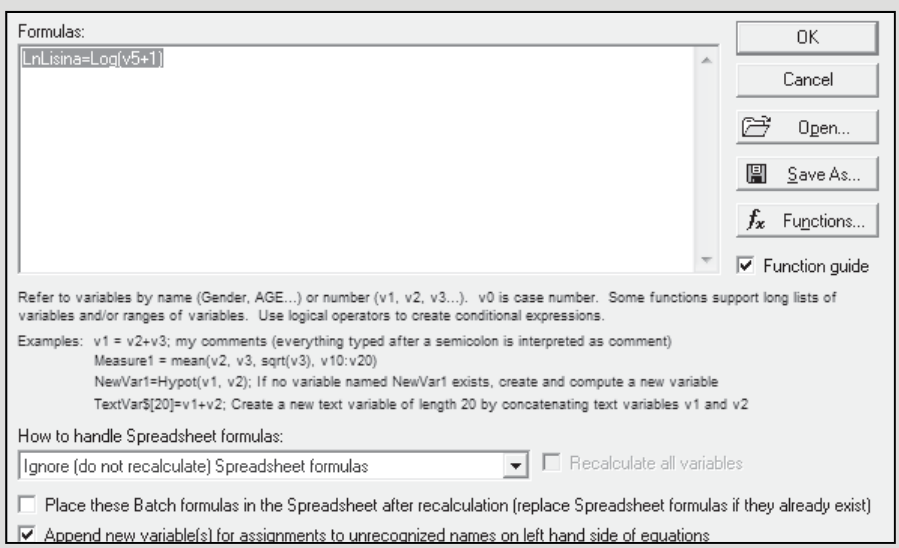


CUADRO IX.3. *(Continuación)* STATISTICA

Paso 13. Al pulsar «Summary: Descriptive statistics» en la anterior ventana aparece la siguiente tabla donde el coeficiente de asimetría es 2,54. En la Figura IX.1 vemos que para esta transformación lo más adecuado podría ser ln(*x*).

	Descriptive Statistics (Spreadsheet3)						
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.	Skewness	Kurtosis
LisiRES1	238	-0,000000	-9,72382	67,41618	9,608332	2,546888	10,89985

Paso 14. Para poder transformar los datos seguimos los pasos 10 y 11 del Cuadro IX.1. En la ventana siguiente introducimos la transformación «LnLisina=Log(v5+1)». Cuando puede haber valores menores o iguales que cero, es necesario sumar el valor 1 u otro mayor, a la variable Lisina (v5) para poder aplicar el logaritmo.


Formulas:
LnLisina=Log(v5+1)
OK
Cancel
Open...
Save As...
Functions...
Function guide
Refer to variables by name (Gender, AGE...) or number (v1, v2, v3...). v0 is case number. Some functions support long lists of variables and/or ranges of variables. Use logical operators to create conditional expressions.
Examples: v1 = v2+v3; my comments (everything typed after a semicolon is interpreted as comment)
Measure1 = mean(v2, v3, sqrt(v3), v10:v20)
NewVar1=Hypot(v1, v2); If no variable named NewVar1 exists, create and compute a new variable
TextVar$[20]=v1+v2; Create a new text variable of length 20 by concatenating text variables v1 and v2
How to handle Spreadsheet formulas:
Ignore (do not recalculate) Spreadsheet formulas
Recalculate all variables
Place these Batch formulas in the Spreadsheet after recalculation (replace Spreadsheet formulas if they already exist)
Append new variable(s) for assignments to unrecognized names on left hand side of equations

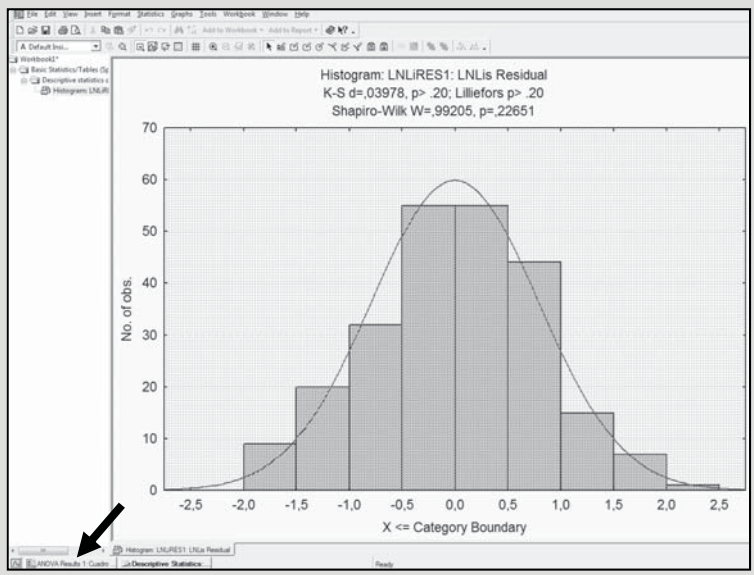
Paso 15. Se repiten los pasos del 1 al 11, pero esta vez usando la variable transformada «LnLisina» en vez de «Lisina».

Como la probabilidad del test de Levene es mayor que 0,05 (p = 0,662), se cumple la homogeneidad de varianzas.

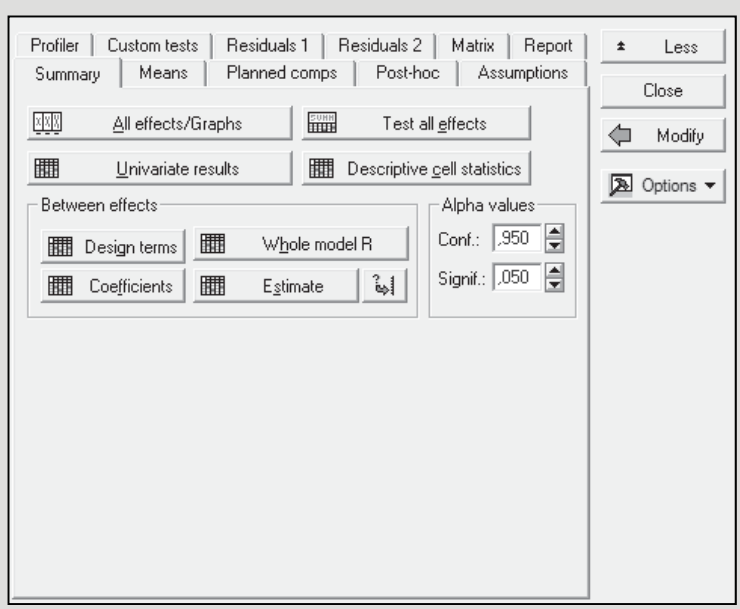
	Levene's Test for Homogeneity of Variances (Cuadro VIII.3.sta) Effect: Especie Degrees of freedom for all F's: 1, 236			
	MS Effect	MS Error	F	p
LnLisina	0,041731	0,217353	0,191996	0,661662

CUADRO IX.3. *(Continuación)*[STATISTICA]

Ahora se observa que, tanto el test Kolmogorov–Smirnov con la corrección de Lilliefors como el de Shapiro-Wilk, muestran valores de $p > 0,2$ y, por tanto, se acepta la hipótesis nula de bondad de ajuste y los residuos se ajustan a una distribución Normal.


Histogram: LNLiRES1: LNLis Residual
K-S d=,03978, p> .20; Lilliefors p> .20
Shapiro-Wilk W=,99205, p=,22651
No. of obs.
X <= Category Boundary

Paso 16. Al pulsar en la pestaña que indica la flecha en la ventana anterior «ANOVA Results», aparece de nuevo la ventana del menú principal del ANOVA, donde en la pestaña «Summary» pulsamos en «Univariate results» para obtener los resultados.



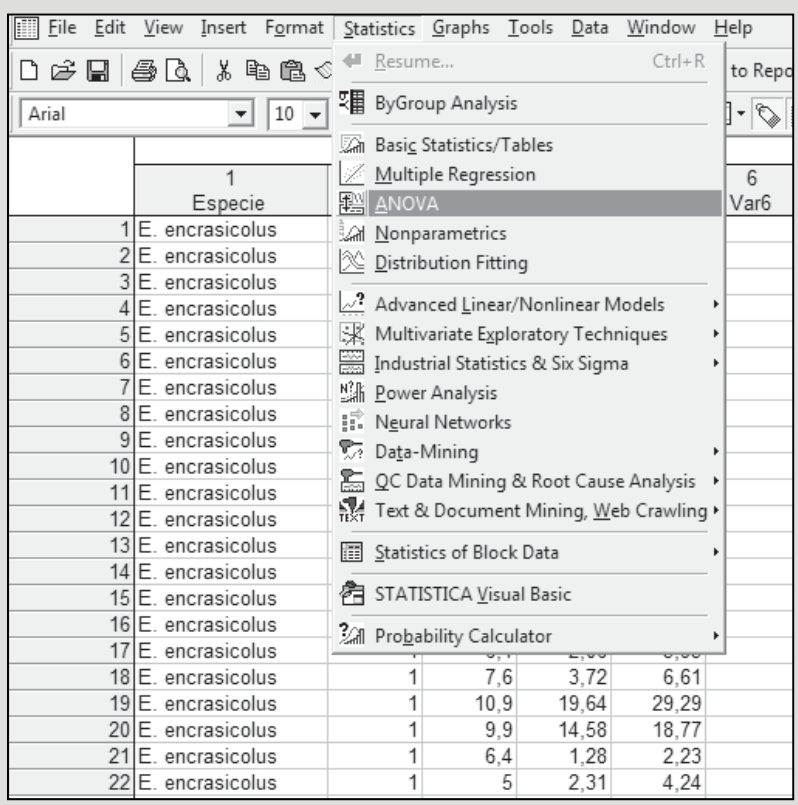
CUADRO IX.3. *(Continuación)*[STATISTICA]

Paso 17. Como $p > 0,05$, se acepta la hipótesis nula del contraste de homogeneidad y, por tanto, se concluye que no existen diferencias significativas en la cantidad del aminoácido Lisina (en pg larva^{-1}), entre las especies *Sardinella aurita* y *Engraulis encrasicolus* (ANOVA, $F_{1, 236} = 0,047$, $p = 0,828$).

	Univariate Results for Each DV (Cuadro VIII.3.sta) Sigma-restricted parameterization Effective hypothesis decomposition				
Effect	Degr. of Freedom	LnLisina SS	LnLisina MS	LnLisina F	LnLisina p
Intercept	1	1022,809	1022,809	1619,552	0,000000
Especie	1	0,030	0,030	0,047	0,828119
Error	236	149,043	0,632		
Total	237	149,073			

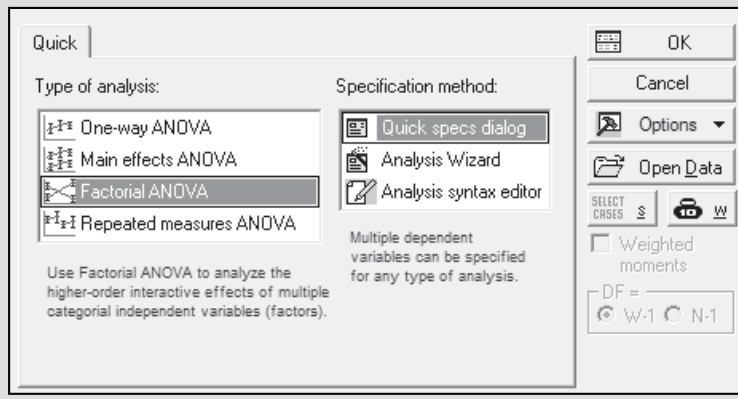
ANOVA multifactorial

Paso 18. Ahora se va a trabajar con los dos factores «Especie» y «Zona». Se selecciona «Statistics» ⇨ «ANOVA».

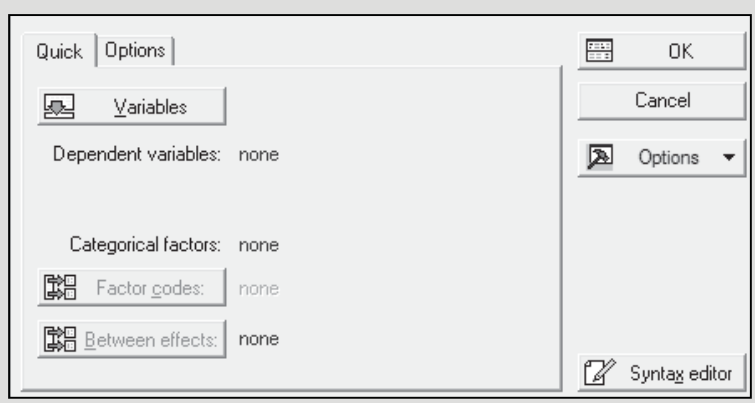


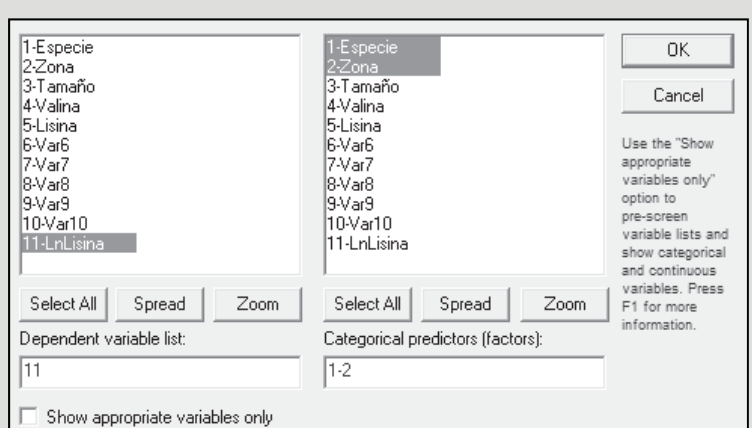
CUADRO IX.3. *(Continuación)*[STATISTICA]

Paso 19. En «Type of analysis» se selecciona «Factorial ANOVA» y en «Specification method:» marcamos «Quick specs dialog».



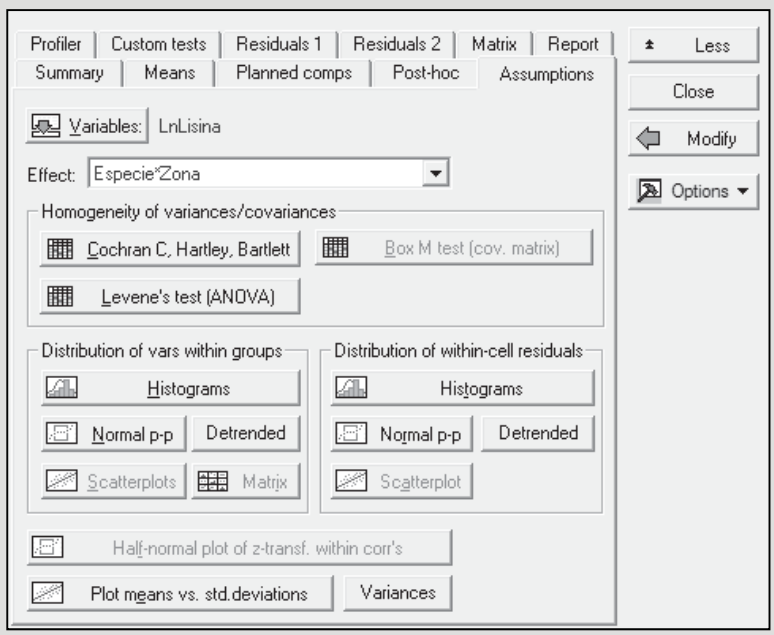
Paso 20. En la siguiente ventana, en la pestaña «Quick» se pulsa en «Variables» y aparece otra ventana donde en «Dependent variable list» introducimos «LnLisina» y en «Categorical predictor (factors):» introducimos los factores «Especie» y «Zona».





CUADRO IX.3. *(Continuación)* STATISTICA

Paso 21. Seguimos los pasos 4 y 5, pero poniendo en «Effect» la interacción «Especie*Zona». Al pulsar «Levene´s test (ANOVA)», obtenemos los resultados del test de Levene. Como la probabilidad es mayor que 0,05, aceptamos la hipótesis nula de homogeneidad de varianzas (p = 0,392).

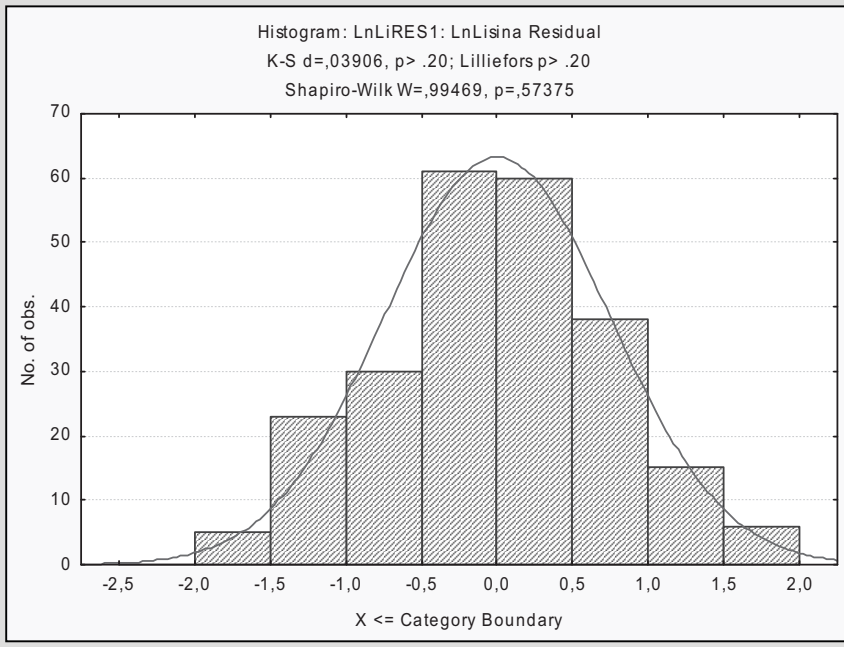


	Levene's Test for Homogeneity of Variances (Cuadro IX.3.sta) Effect: Especie*Zona Degrees of freedom for all F's: 5, 232			
	MS Effect	MS Error	F	p
LnLisina	0,221353	0,212179	1,043238	0,392945

Paso 22. Seguimos los pasos 6 a 11. Se obtiene la tabla de residuos y luego el siguiente gráfico donde se observa que, tanto el test Kolmogorov–Smirnov con la corrección de Lilliefors como el de Shapiro-Wilk, muestran valores de p > 0,2 y, por tanto, se acepta la hipótesis nula de bondad de ajuste y los residuos se ajustan a una distribución Normal.

	Cuadro IX.3.sta		
	1 LnLiOBS1	2 LnLiPRE1	3 LnLiRES1
1	2,819	2,382	0,437
2	1,744	2,382	-0,638
3	1,959	2,382	-0,423
4	3,419	2,382	1,038
5	0,615	2,382	-1,766
6	3,741	2,382	1,359

CUADRO IX.3. *(Continuación)* STATISTICA

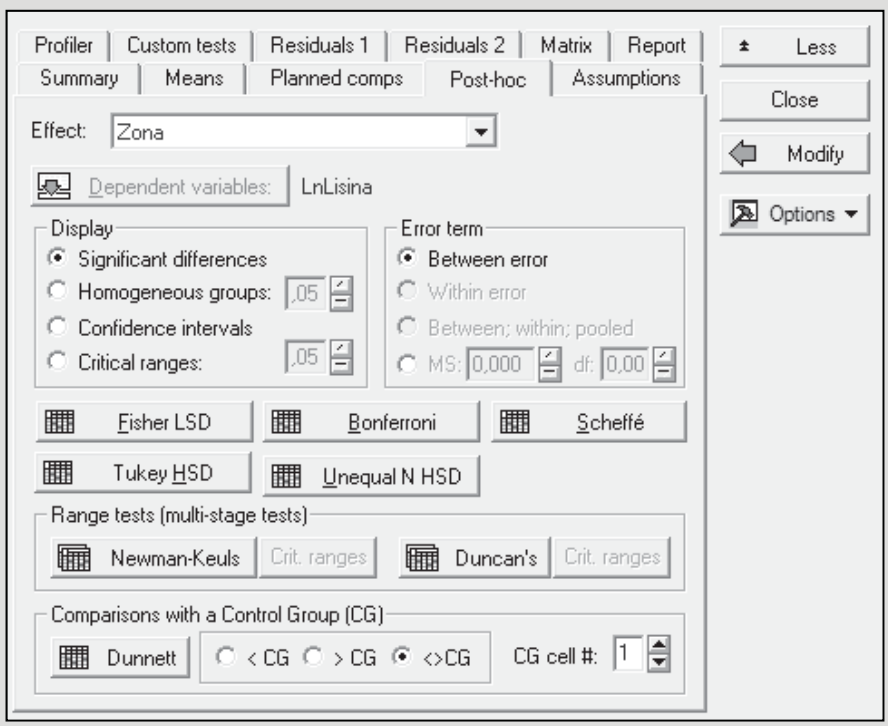


Paso 23. Repetimos el paso 16 y se obtiene la tabla de ANOVA. Los resultados se muestran en la siguiente tabla. Es interesante resaltar que, a pesar de considerar el efecto zona, se siguen sin observar diferencias significativas en la cantidad del aminoácido Lisina (en pg larva^{-1}) entre las especies (ANOVA, $F_{1,\ 232}$ = 1,53, p = 0,217). Sin embargo, existen diferencias significativas entre zonas (ANOVA, $F_{2,\ 232}$ = 9,57, p < 0,001), En la tabla de ANOVA se muestra también el efecto conjunto de ambos factores, «Especie*Zona», considerado como un factor adicional que recoge la posibilidad de que –además del efecto de la especie y de la zona, que se suman– aparezca otro más, que también se suma a los anteriores, que es la interacción de ambos. En este ejemplo la interacción no es significativa (ANOVA, $F_{2,\ 232}$ = 1,6, p = 0,192).

	Univariate Results for Each DV (Cuadro IX.3.sta) Sigma-restricted parameterization Effective hypothesis decomposition				
Effect	Degr. of Freedom	LnLisina SS	LnLisina MS	LnLisina F	LnLisina p
Intercept	1	870,0202	870,0202	1510,790	0,000000
Especie	1	0,8813	0,8813	1,530	0,217307
Zona	2	11,0184	5,5092	9,567	0,000102
Especie*Zona	2	1,9142	0,9571	1,662	0,192008
Error	232	133,6020	0,5759		
Total	237	149,0728			

CUADRO IX.3. *(Continuación)*[STATISTICA]

Paso 24. Como el factor «Zona» tiene más de 2 niveles (hay tres zonas), es posible hacer pruebas *post hoc* para ver las diferencias específicas que existen entre cada par de zonas. Para ello en la ventana del menú principal del ANOVA seleccionamos la pestaña «Post-hoc» y aparece la siguiente ventana. En «Effect» se introduce el factor «Zona». Para varianzas iguales, que es nuestro caso, uno que se utiliza frecuentemente es el test Tukey, por tanto pulsamos en «Tukey HSD».



Paso 25. Se observa que hay diferencias entre la zona 1 y la 3 (p < 0,001) y entre la 2 y la 3 (p = 0,016), mientras que no existen diferencias significativas en la cantidad del aminoácido Lisina (en pg larva^{-1}) entre las zonas 1 y 2 (p = 0,094).

	Tukey HSD test; variable LnLisina (Cuadro IX.3.sta) Approximate Probabilities for Post Hoc Tests Error: Between MS = ,57587, df = 232,00			
Cell No.	Zona	{1} 2,4016	{2} 2,1406	{3} 1,8242
1	1		0,094092	0,000035
2	2	0,094092		0,015717
3	3	0,000035	0,015717	

CUADRO IX.3. *(Continuación)*[SPSS]

Aplicación con SPSS.

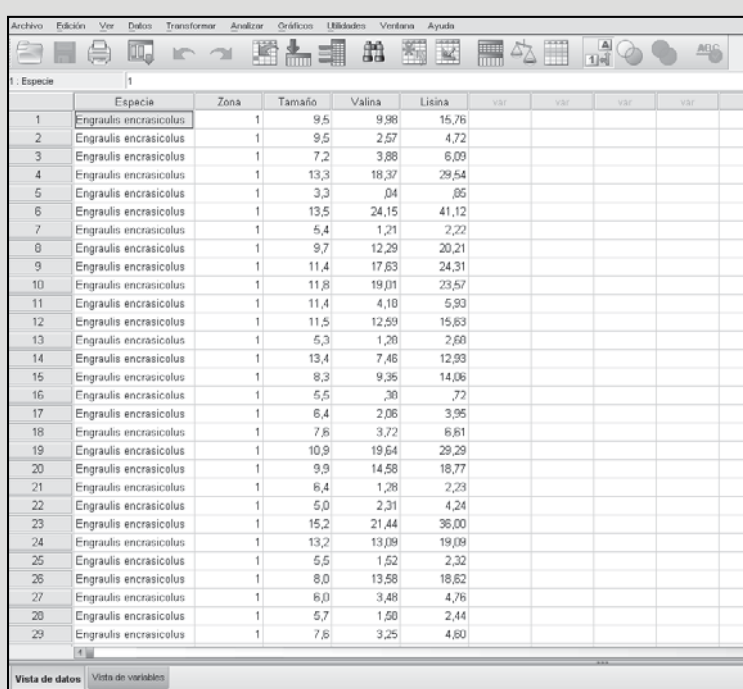
EJEMPLO. Vamos a utilizar los mismos datos que en la aplicación con STATISTICA, es decir, datos del tamaño de las larvas de dos especies de peces de la familia Clupeidae, *Sardinella aurita* y *Engraulis encrasicolus*, y la cantidad de dos aminoácidos (Valina y Lisina) en pg larva^{-1}, en diferentes zonas del Canal de Sicilia. Los datos están en el archivo **Cuadro IX.3.sav**.

En este caso, el objetivo es determinar si existen diferencias significativas en la cantidad del aminoácido Valina entre las especies y/o las zonas.

Para abordar el objetivo planteado es necesario hacer un ANOVA multifactorial, es decir, considerar los dos factores («Especie» y «Zona»). Sin embargo, para explicar como se hace un ANOVA con un factor (unifactorial), en primer lugar vamos a trabajar solo con el factor «Especie» y luego realizaremos el análisis considerando los dos factores («Especie» y «Zona»). Si se quiere hacer directamente un ANOVA multifactorial, éste se describe a partir del paso 23.

ANOVA unifactorial

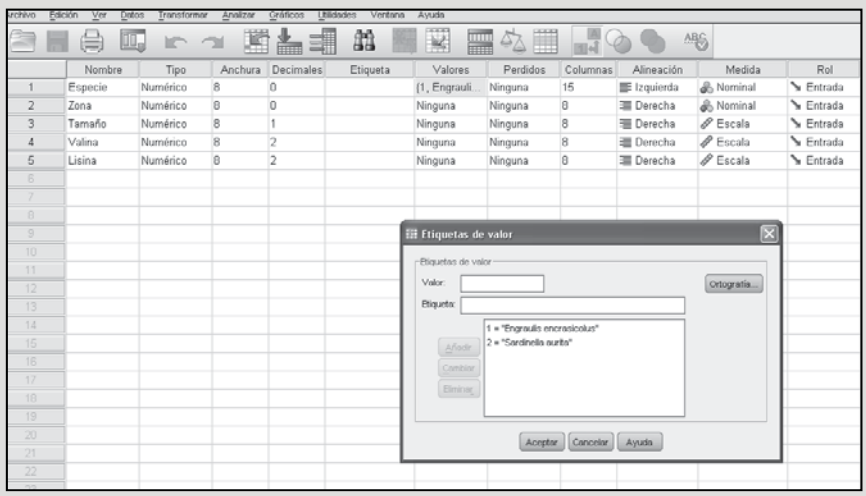
Paso 1. En SPSS los datos hay que organizarlos como se muestra en el archivo **Cuadro IX.3.sav** y en la siguiente ventana.


Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana Ayuda
1 : Especie 1
Vista de datos
Vista de variables

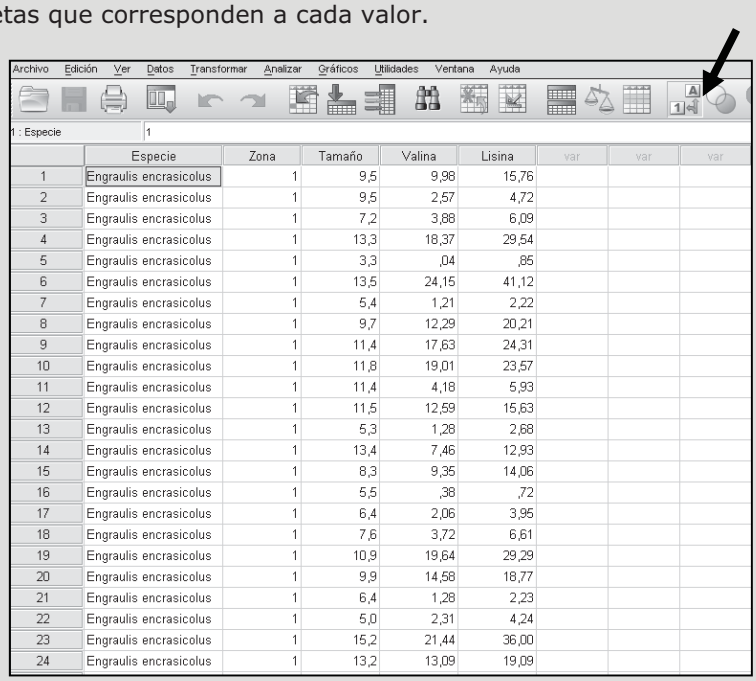
	Especie	Zona	Tamaño	Valina	Lisina
1	Engraulis encrasicolus	1	9,5	9,98	15,76
2	Engraulis encrasicolus	1	9,5	2,57	4,72
3	Engraulis encrasicolus	1	7,2	3,88	6,09
4	Engraulis encrasicolus	1	13,3	18,37	29,54
5	Engraulis encrasicolus	1	3,3	,04	,85
6	Engraulis encrasicolus	1	13,5	24,15	41,12
7	Engraulis encrasicolus	1	5,4	1,21	2,22
8	Engraulis encrasicolus	1	9,7	12,29	20,21
9	Engraulis encrasicolus	1	11,4	17,63	24,31
10	Engraulis encrasicolus	1	11,8	19,01	23,57
11	Engraulis encrasicolus	1	11,4	4,10	5,93
12	Engraulis encrasicolus	1	11,5	12,59	15,63
13	Engraulis encrasicolus	1	5,3	1,26	2,60
14	Engraulis encrasicolus	1	13,4	7,46	12,93
15	Engraulis encrasicolus	1	8,3	9,36	14,06
16	Engraulis encrasicolus	1	5,5	,38	,72
17	Engraulis encrasicolus	1	6,4	2,06	3,95
18	Engraulis encrasicolus	1	7,6	3,72	6,61
19	Engraulis encrasicolus	1	10,9	19,64	29,29
20	Engraulis encrasicolus	1	9,9	14,58	18,77
21	Engraulis encrasicolus	1	6,4	1,28	2,23
22	Engraulis encrasicolus	1	5,0	2,31	4,24
23	Engraulis encrasicolus	1	15,2	21,44	36,00
24	Engraulis encrasicolus	1	13,2	13,09	19,09
25	Engraulis encrasicolus	1	5,5	1,52	2,32
26	Engraulis encrasicolus	1	8,0	13,58	18,62
27	Engraulis encrasicolus	1	6,0	3,48	4,76
28	Engraulis encrasicolus	1	5,7	1,50	2,44
29	Engraulis encrasicolus	1	7,6	3,25	4,60

CUADRO IX.3. *(Continuación SPSS*

Paso 2. Al seleccionar la pestaña «Vista de variables» aparece la siguiente ventana donde es posible especificar el nombre de cada variable en «Nombre», el tipo de variable en «Tipo», etc. En «Valores» aparece la ventana que se muestra a continuación, donde es posible etiquetar a qué corresponde cada valor en cada una de las variables. Se muestra el ejemplo para la variable «Especie».


Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana Ayuda
Nombre Tipo Anchura Decimales Etiqueta Valores Perdidos Columnas Alineación Medida Rol
1 Especie Numérico 8 0 (1, Engraulí... Ninguna 15 Izquierda Nominal Entrada
2 Zona Numérico 8 0 Ninguna Ninguna 8 Derecha Nominal Entrada
3 Tamaño Numérico 8 1 Ninguna Ninguna 8 Derecha Escala Entrada
4 Valina Numérico 8 2 Ninguna Ninguna 8 Derecha Escala Entrada
5 Lisina Numérico 8 2 Ninguna Ninguna 8 Derecha Escala Entrada
Etiquetas de valor
Etiquetas de valor
Valor:
Etiqueta:
Ortografía...
1 = "Engraulis encrasicolus"
2 = "Sardinella aurita"
Añadir
Cambiar
Eliminar
Aceptar Cancelar Ayuda

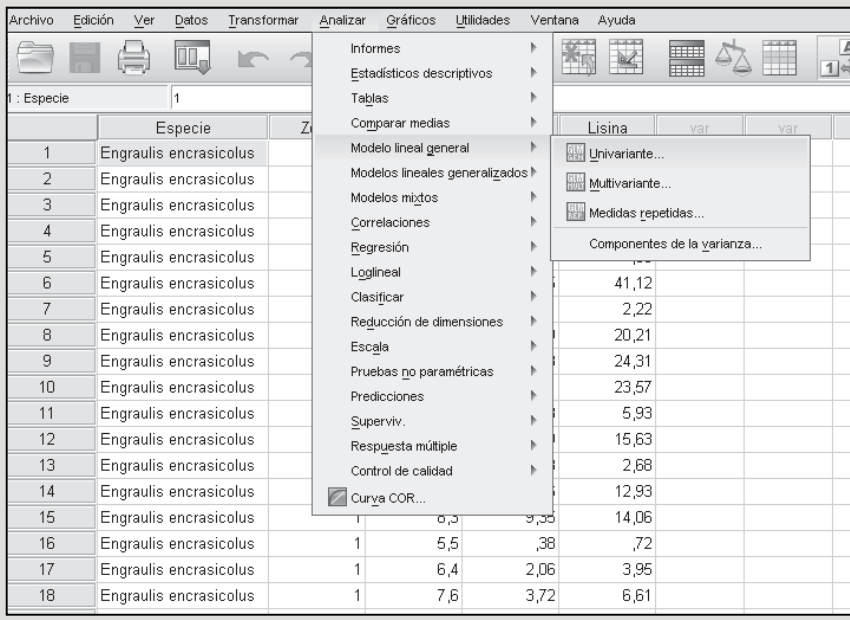
Paso 3. Al volver a la ventana de la pestaña «Vista de datos», si pulsamos donde indica la flecha, aparecen en vez de los números las etiquetas que corresponden a cada valor.



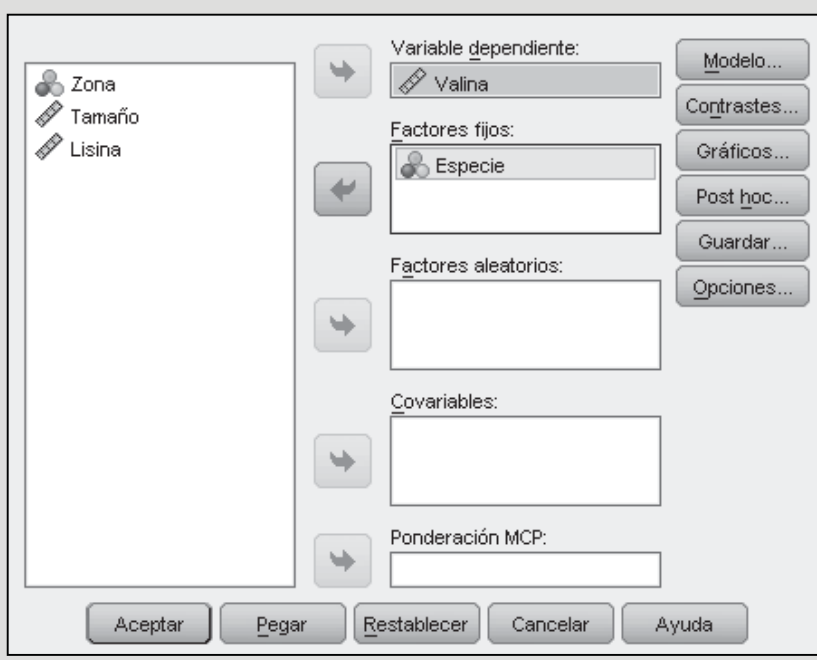
	Especie	Zona	Tamaño	Valina	Lisina	var	var	var
1	Engraulis encrasicolus	1	9,5	9,98	15,76			
2	Engraulis encrasicolus	1	9,5	2,57	4,72			
3	Engraulis encrasicolus	1	7,2	3,88	6,09			
4	Engraulis encrasicolus	1	13,3	18,37	29,54			
5	Engraulis encrasicolus	1	3,3	,04	,85			
6	Engraulis encrasicolus	1	13,5	24,15	41,12			
7	Engraulis encrasicolus	1	5,4	1,21	2,22			
8	Engraulis encrasicolus	1	9,7	12,29	20,21			
9	Engraulis encrasicolus	1	11,4	17,63	24,31			
10	Engraulis encrasicolus	1	11,8	19,01	23,57			
11	Engraulis encrasicolus	1	11,4	4,18	5,93			
12	Engraulis encrasicolus	1	11,5	12,59	15,63			
13	Engraulis encrasicolus	1	5,3	1,28	2,68			
14	Engraulis encrasicolus	1	13,4	7,46	12,93			
15	Engraulis encrasicolus	1	8,3	9,35	14,06			
16	Engraulis encrasicolus	1	5,5	,38	,72			
17	Engraulis encrasicolus	1	6,4	2,06	3,95			
18	Engraulis encrasicolus	1	7,6	3,72	6,61			
19	Engraulis encrasicolus	1	10,9	19,64	29,29			
20	Engraulis encrasicolus	1	9,9	14,58	18,77			
21	Engraulis encrasicolus	1	6,4	1,28	2,23			
22	Engraulis encrasicolus	1	5,0	2,31	4,24			
23	Engraulis encrasicolus	1	15,2	21,44	36,00			
24	Engraulis encrasicolus	1	13,2	13,09	19,09			

CUADRO IX.3. *(Continuación)*[SPSS]

Paso 4. Para realizar el ANOVA unifactorial vamos a «Analizar» ⇨ «Modelo lineal general» ⇨ «Univariante…».



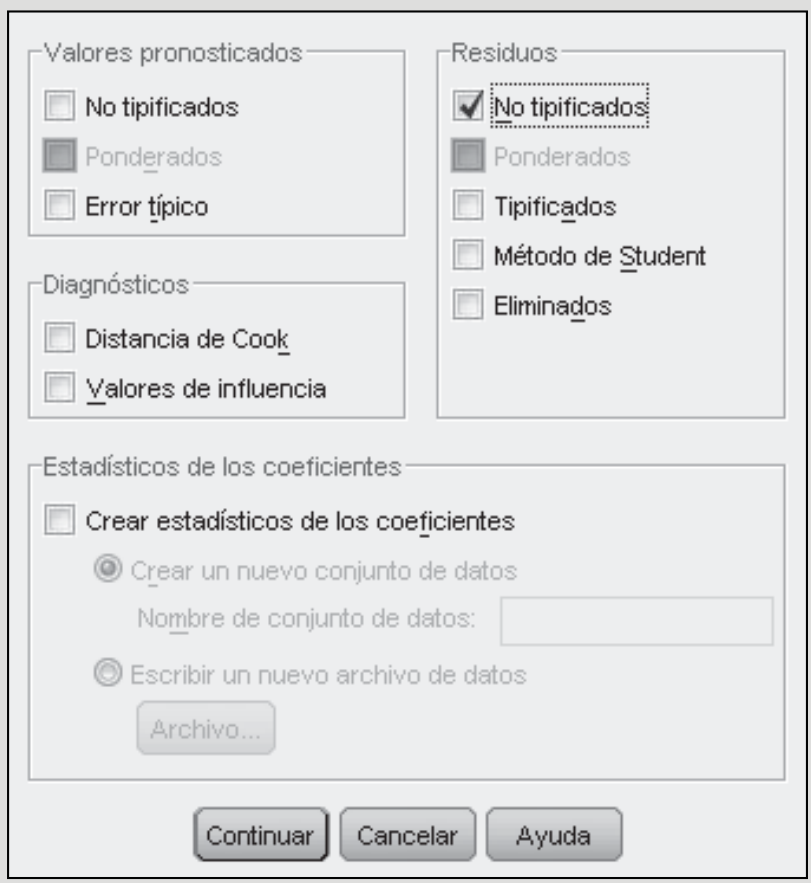
Paso 5. En «Variable dependiente» se incluye la variable a analizar, en este caso «Valina». En «Factores fijos» se incluye «Especie».



CUADRO IX.3. *(Continuación)*[SPSS]

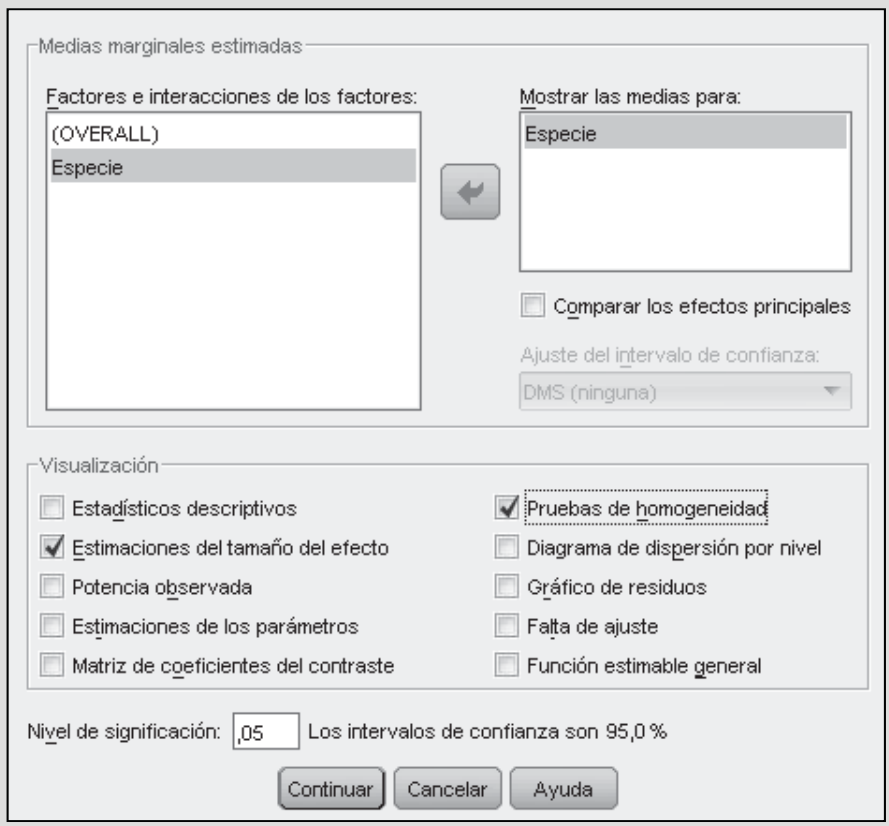
Como en el factor solo hay dos niveles, es decir dos especies, no es necesario realizar «Contrastes» ni pruebas «Post hoc», las cuales se usan cuando hay más de dos niveles en los factores y permiten comparar las medias entre cada par de niveles dentro de cada factor. Más adelante cuando se utilice el factor «Zona», se explicaran los «Contrastes» y las pruebas «Post hoc».

Paso 6. En la ventana anterior en «Guardar…» aparece la siguiente ventana, donde en «Residuos» se marca «No tipificados...» para poder testar la Normalidad de los residuos.



Paso 7. En «Opciones» se marca «Estimaciones del tamaño del efecto», que permite cuantificar la importancia del factor sobre la variable dependiente por medio del estadístico Eta al cuadrado parcial, y «Pruebas de homogeneidad» para ver la homogeneidad de varianzas.

CUADRO IX.3. *(Continuación)*[SPSS]



Paso 8. El test de Levene muestra que se cumple el requisito de homogeneidad de varianzas, ya que la probabilidad es mayor que 0,05 (p = 0,489).

Contraste de Levene sobre la igualdad de las varianzas error[a]

Variable dependiente:Valina

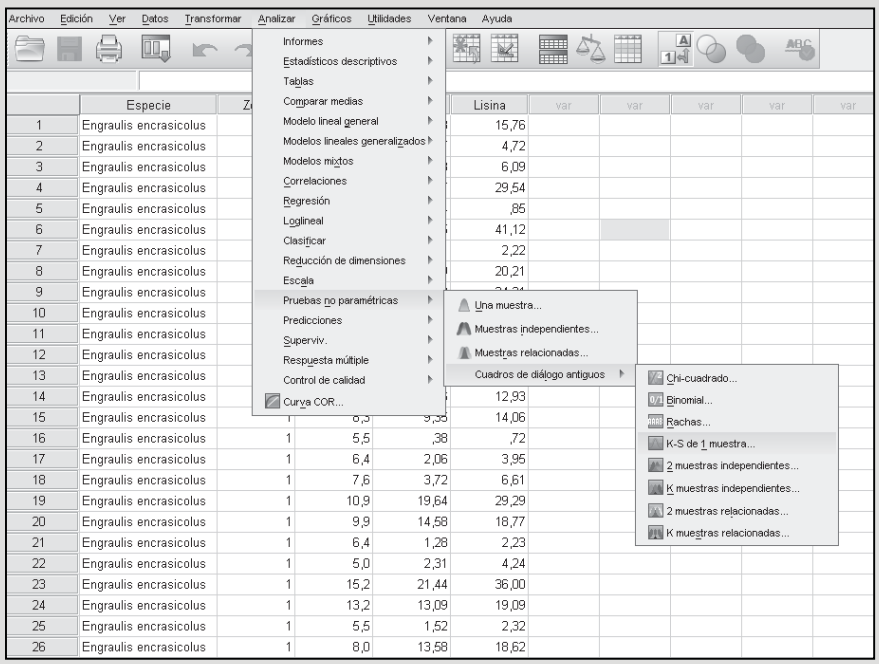
F	gl1	gl2	Sig.
,480	1	236	,489

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

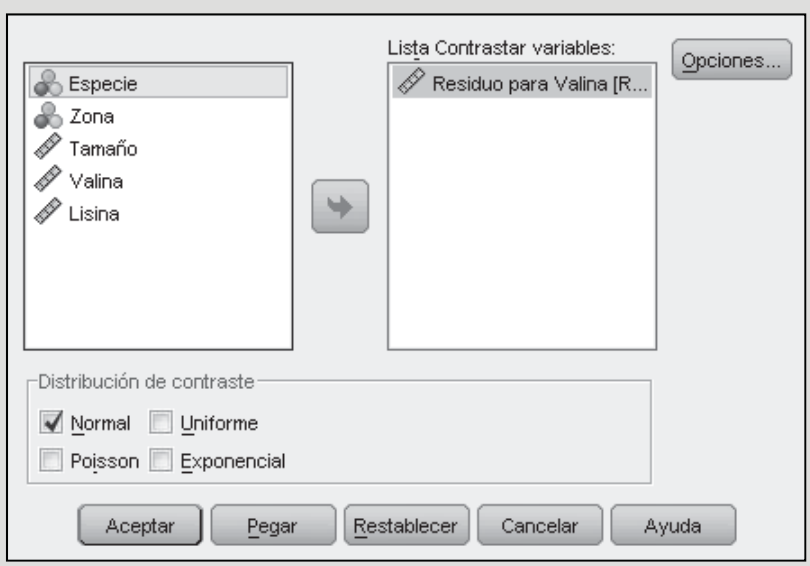
a. Diseño: Intersección + Especie

Paso 9. Antes de comenzar a examinar los resultados del ANOVA es necesario comprobar el otro requisito, es decir, si los residuos tienen distribución Normal. Para ello vamos a «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «K-S de una muestra…».

CUADRO IX.3. *(Continuación)*[SPSS]



Paso 10. En «Lista Contrastar variables:» se introducen los residuos y en «Distribución de contraste» se marca «Normal».

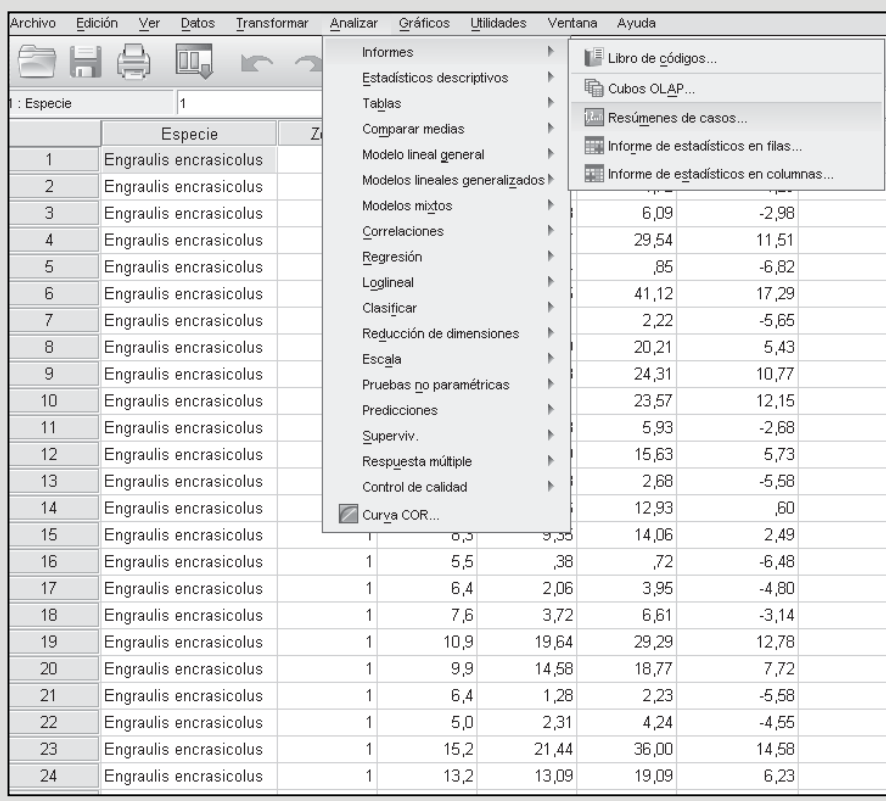


Paso 11. El test de Kolmogorov-Smirnov muestra que la distribución de los residuos no es Normal, ya que $p < 0{,}001$. Por tanto, es necesario transformar los datos.

CUADRO IX.3. *(Continuación)*[SPSS]

		Residuo para Valina
N		238
Parámetros normales[a,,b]	Media	,0000
	Desviación típica	6,66542
Diferencias más extremas	Absoluta	,189
	Positiva	,189
	Negativa	-,149
Z de Kolmogorov-Smirnov		2,920
Sig. asintót. (bilateral)		,000

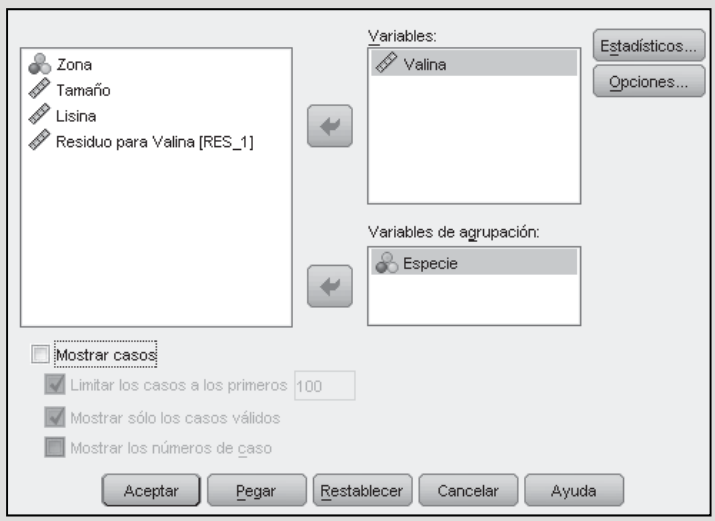
Paso 12. Para saber cual es la mejor transformación es conveniente conocer la asimetría de los datos. Para ello se va a «Analizar» ⇨ «Informes» ⇨ «Resúmenes de casos…».



	Especie					
1	Engraulis encrasicolus				6,09	-2,98
2	Engraulis encrasicolus				29,54	11,51
3	Engraulis encrasicolus				,85	-6,82
4	Engraulis encrasicolus				41,12	17,29
5	Engraulis encrasicolus				2,22	-5,65
6	Engraulis encrasicolus				20,21	5,43
7	Engraulis encrasicolus				24,31	10,77
8	Engraulis encrasicolus				23,57	12,15
9	Engraulis encrasicolus				5,93	-2,68
10	Engraulis encrasicolus				15,63	5,73
11	Engraulis encrasicolus				2,68	-5,58
12	Engraulis encrasicolus				12,93	,60
13	Engraulis encrasicolus				14,06	2,49
14	Engraulis encrasicolus					
15	Engraulis encrasicolus					
16	Engraulis encrasicolus	1	5,5	,38	,72	-6,48
17	Engraulis encrasicolus	1	6,4	2,06	3,95	-4,80
18	Engraulis encrasicolus	1	7,6	3,72	6,61	-3,14
19	Engraulis encrasicolus	1	10,9	19,64	29,29	12,78
20	Engraulis encrasicolus	1	9,9	14,58	18,77	7,72
21	Engraulis encrasicolus	1	6,4	1,28	2,23	-5,58
22	Engraulis encrasicolus	1	5,0	2,31	4,24	-4,55
23	Engraulis encrasicolus	1	15,2	21,44	36,00	14,58
24	Engraulis encrasicolus	1	13,2	13,09	19,09	6,23

CUADRO IX.3. *(Continuación)*[SPSS]

Paso 13. En «Variables» se introduce «Valina» y en «Variables de agrupación:» se incluye «Especie». No se marca «Mostrar casos».



Paso 14. Al pulsar «Estadísticos» en la ventana anterior aparece la siguiente ventana donde en «Estadísticos de casilla:» se introduce «Asimetría».

CUADRO IX.3. *(Continuación)*[SPSS]

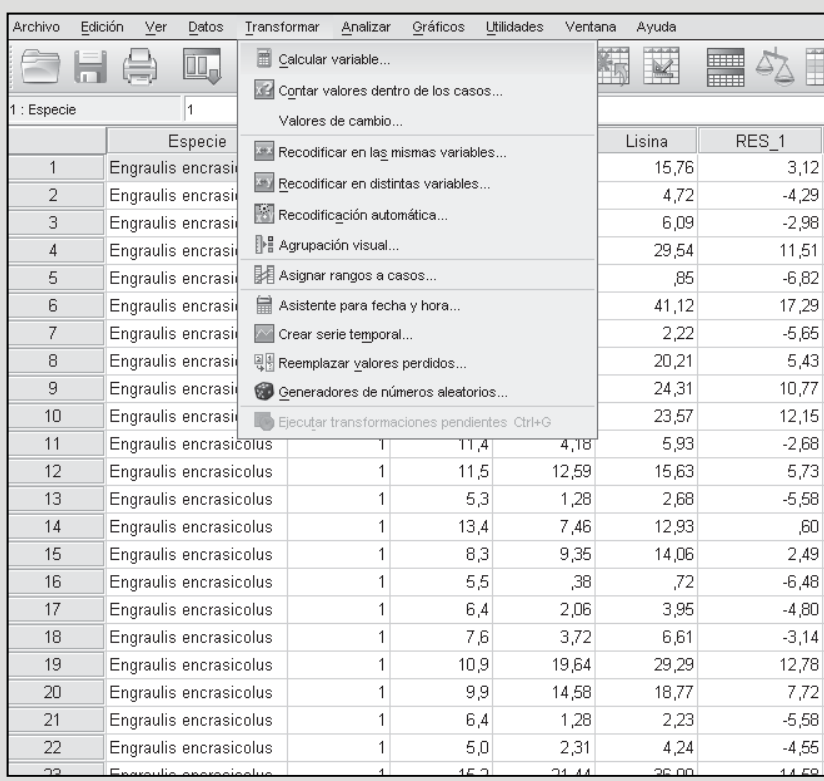
Paso 15. Como la asimetría es positiva y relativamente alta, en la Figura IX.1 vemos que en este caso lo más adecuado podría ser la transformación ln(x).

Resúmenes de casos

Valina

Especie	N	Asimetría
Engraulis encrasicolus	136	1,512
Sardinella aurita	102	3,074
Total	238	2,437

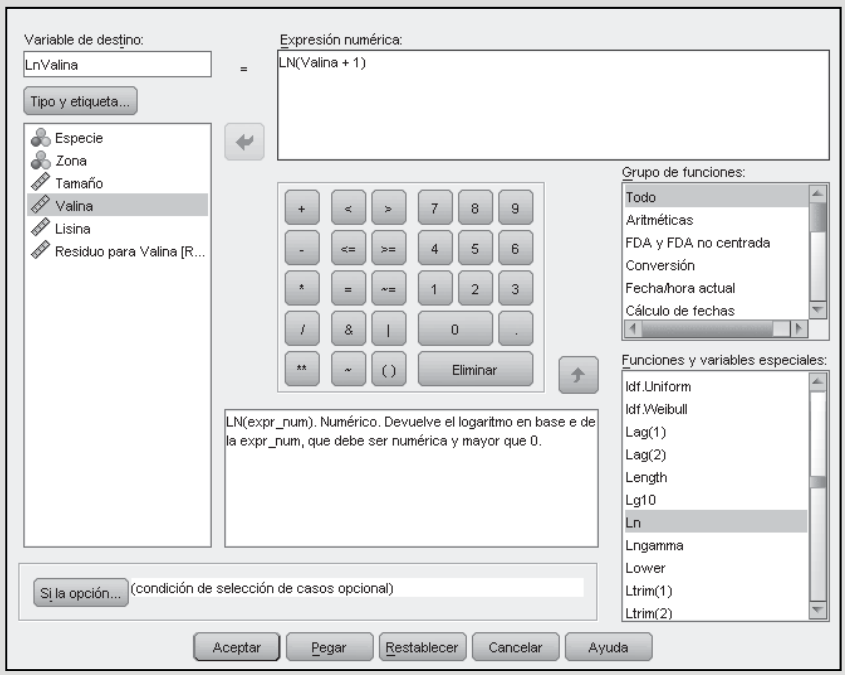
Paso 16. Para realizar la transformación vamos a «Transformar» ⇨ «Calcular variable…».



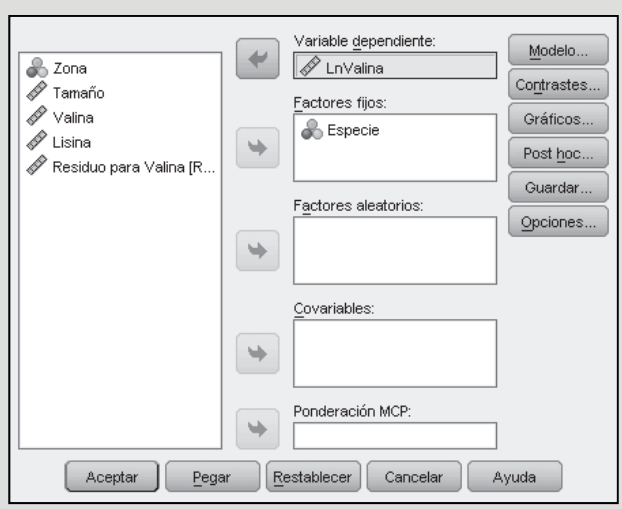
	Especie				Lisina	RES_1
1	Engraulis encrasi				15,76	3,12
2	Engraulis encrasi				4,72	-4,29
3	Engraulis encrasi				6,09	-2,98
4	Engraulis encrasi				29,54	11,51
5	Engraulis encrasi				,85	-6,82
6	Engraulis encrasi				41,12	17,29
7	Engraulis encrasi				2,22	-5,65
8	Engraulis encrasi				20,21	5,43
9	Engraulis encrasi				24,31	10,77
10	Engraulis encrasi				23,57	12,15
11	Engraulis encrasicolus	1	11,4	4,18	5,93	-2,68
12	Engraulis encrasicolus	1	11,5	12,59	15,63	5,73
13	Engraulis encrasicolus	1	5,3	1,28	2,68	-5,58
14	Engraulis encrasicolus	1	13,4	7,46	12,93	,60
15	Engraulis encrasicolus	1	8,3	9,35	14,06	2,49
16	Engraulis encrasicolus	1	5,5	,38	,72	-6,48
17	Engraulis encrasicolus	1	6,4	2,06	3,95	-4,80
18	Engraulis encrasicolus	1	7,6	3,72	6,61	-3,14
19	Engraulis encrasicolus	1	10,9	19,64	29,29	12,78
20	Engraulis encrasicolus	1	9,9	14,58	18,77	7,72
21	Engraulis encrasicolus	1	6,4	1,28	2,23	-5,58
22	Engraulis encrasicolus	1	5,0	2,31	4,24	-4,55

Paso 17. En «Variable de destino:» creamos una nueva variable «LnValina» y en «Expresión numérica:» introducimos la transformación «ln(Valina+1)». Como hay valores que son menores o iguales que cero, para poder aplicar mejor la transformación es necesario sumar 1 a la variable.

CUADRO IX.3. *(Continuación)*[SPSS]



Paso 18. Se repiten los pasos del 4 al 11 para realizar el ANOVA, pero en este caso se trabaja con la variable dependiente «LnValina» en vez de «Valina».



Paso 19. El último paso antes de ver los resultados del ANOVA es comprobar si se cumplen los requisitos de Normalidad y homogeneidad de varianzas. Para ello seleccionamos «Analizar» ⇨ «Estadísticos descriptivos» ⇨ «Explorar...».

CUADRO IX.3. *(Continuación)*[SPSS]

Paso 20. Con los datos transformados, el test de Kolmogorov-Smirnov muestra que la distribución de los residuos es Normal, ya que la probabilidad es mayor que 0,05 (p = 0,279).

Prueba de Kolmogorov-Smirnov para una muestra

		Residuo para LnValina
N		238
Parámetros normales[a,,b]	Media	,0000
	Desviación típica	,77000
Diferencias más extremas	Absoluta	,064
	Positiva	,064
	Negativa	-,053
Z de Kolmogorov-Smirnov		,992
Sig. asintót. (bilateral)		,279

Paso 21. Los resultados del test de Levene muestran que existe homogeneidad de varianzas (p = 0,963).

Contraste de Levene sobre la igualdad de las varianzas error[a]

Variable dependiente:LnValina

F	gl1	gl2	Sig.
,002	1	236	,963

Paso 22. En los resultados del ANOVA como p > 0,05, se acepta la hipótesis nula de igualdad de medias y, por tanto, se concluye que no existen diferencias significativas en la cantidad del aminoácido Valina (en pg larva^{-1}), entre las especies *Sardinella aurita* y *Engraulis encrasicolus* (ANOVA, $F_{1,\ 236}$ = 0,01, p = 0,97). El estadístico Eta cuadrado parcial informa sobre el porcentaje de la variabilidad residual de la variable dependiente (no explicada ya por otras variables) que es explicada por el factor, en un rango de 0 a 1. Para el factor «Especie» el valor es 0 y, por tanto, este factor «Especie» no es importante a la hora de explicar la variación observada en el aminoácido Valina.

Pruebas de los efectos inter-sujetos

Variable dependiente:LnValina

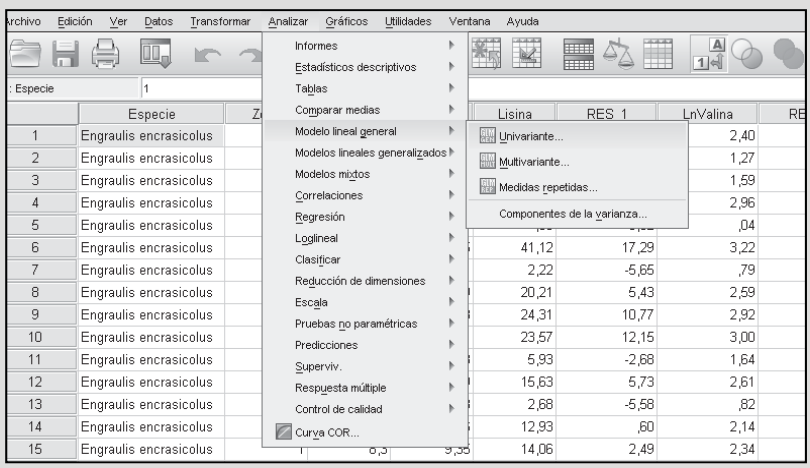
Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial
Modelo corregido	,001[a]	1	,001	,001	,970	,000
Intersección	741,827	1	741,827	1245,889	,000	,841
Especie	,001	1	,001	,001	,970	,000
Error	140,519	236	,595			
Total	898,032	238				
Total corregida	140,520	237				

CUADRO IX.3. *(Continuación)*[SPSS]

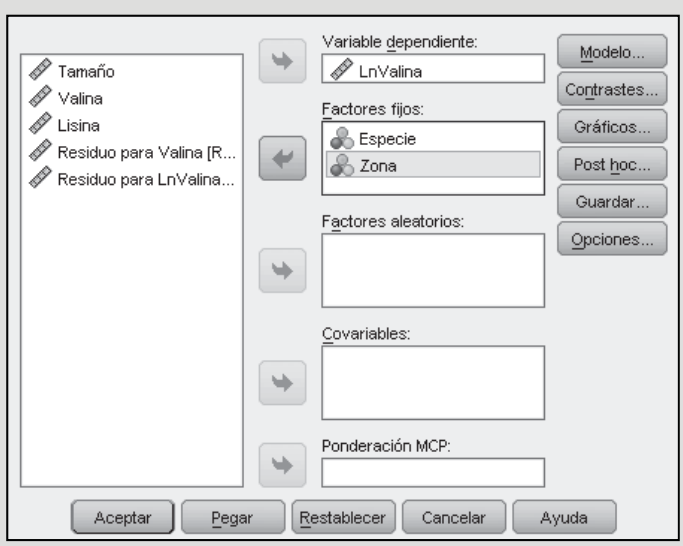
ANOVA multifactorial

A continuación vamos a realizar el análisis considerando los dos factores, «Especie» y «Zona».

Paso 23. Para realizar el análisis se selecciona «Analizar» ⇨ «Modelo lineal general» ⇨ «Univariante…». En el supuesto de querer analizar más de una variable dependiente, simplemente se seleccionaría «Multivariante…».



Paso 24. En «Variable dependiente» se incluye la variable a analizar, en este caso «LnValina». En «Factores fijos» se incluyen «Especie» y «Zona». Se utiliza la variable transformada ya que, en el análisis unifactorial con el factor «Especie» (véase pasos 1 a 22), se observó que era necesario transformar la variable dependiente para cumplir los requisitos de Normalidad y homogeneidad de varianzas.

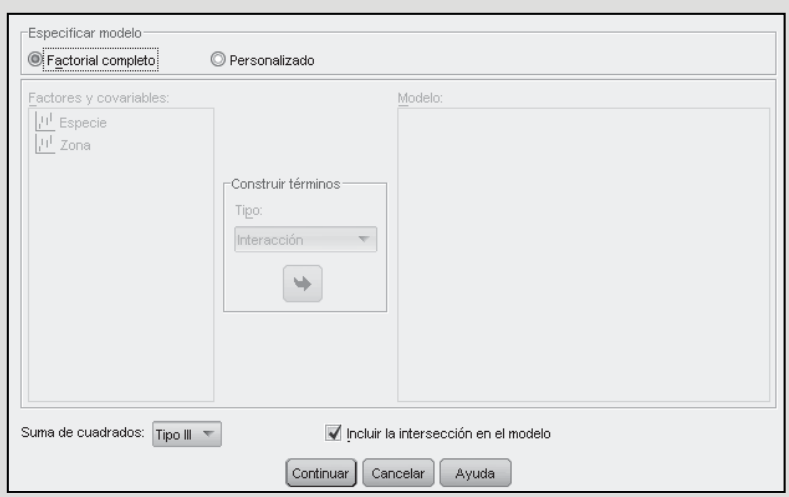


CUADRO IX.3. *(Continuación)*[SPSS]

La diferencia entre los factores fijos y los aleatorios es que en estos últimos los niveles del factor no son todos los existentes, sino una muestra de niveles, lo que añade una fuente de variabilidad adicional. Por ejemplo, imaginemos que consideramos «Zona» como factor aleatorio, eso significa que se consideran solo unas cuantas áreas al azar y no se considerarían todas las áreas de muestreo de las que existe información.

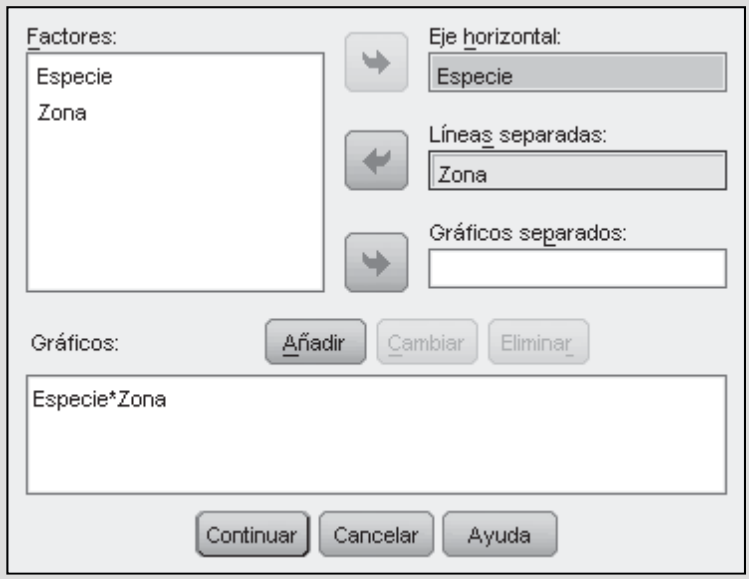
Paso 25. En «Modelo…» se puede especificar el modelo deseado. El «factorial completo», que contiene todos los efectos principales (de cada factor) y todas las interacciones (entre varios factores) y el «personalizado», que sirve para especificar sólo un subconjunto de efectos. El modelo depende de la naturaleza de los datos y del tamaño de la muestra. Nosotros dejaremos el que da por defecto el programa, que es el factorial completo. La muestra puede ser insuficiente para estimar el modelo completo, en cuyo caso suprimiríamos las interacciones de mayor nivel. También debemos suprimir cualquier interacción que no tenga interés en el análisis o que por consideraciones teóricas no debe existir.

En «Suma de cuadrados» se especifica la forma de calcular las sumas de cuadrados. Se usa generalmente el Tipo III porque es adecuado para cualquier modelo en el que no haya celdas vacías, incluyendo aquellas situaciones en las que son adecuados los tipos I y II. El tipo IV es el más conveniente cuando hay celdas vacías.

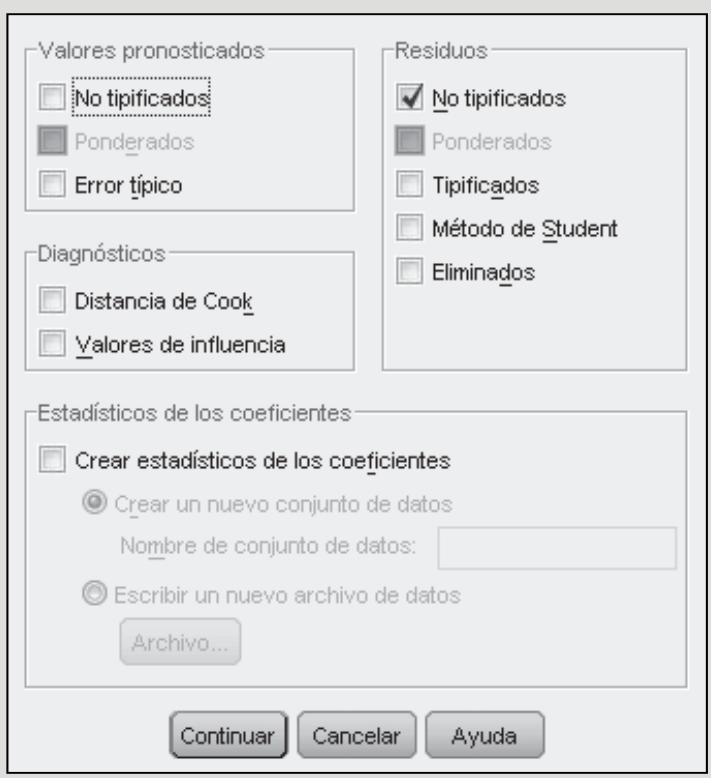


Paso 26. En la opción «Gráficos…» aparecen los gráficos de perfil (gráficos de interacción), que sirven para comparar las medias marginales en el modelo (aumentan o disminuyen). Un gráfico de perfil es un gráfico de líneas en el que cada punto indica la media marginal estimada de una variable dependiente en un nivel de un factor. Todos los factores fijos y aleatorios, si existen, están disponibles para los gráficos. Solo hay que seleccionar los factores que se quieren representar y pulsar «Añadir» para incluir los diferentes gráficos.

CUADRO IX.3. *(Continuación)*[SPSS]

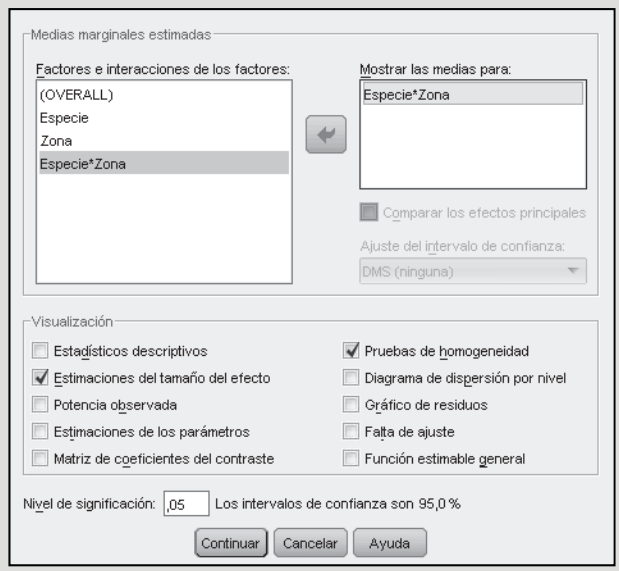


Paso 27. En la opción «Guardar» se marcan los residuos «No tipificados» para más tarde analizar si tienen una distribución Normal.

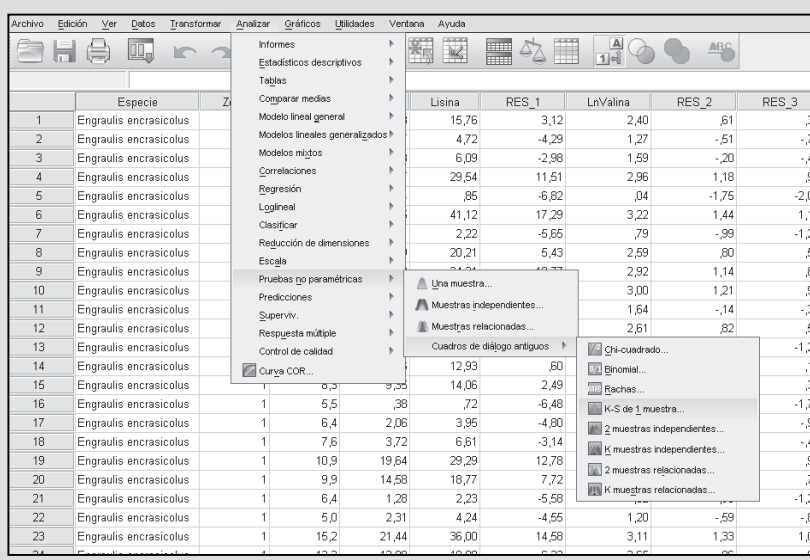


CUADRO IX.3. *(Continuación)*[SPSS]

Paso 28. En «Opciones» se marca «Estimaciones del tamaño del efecto», que permite cuantificar la importancia de cada factor sobre la variable dependiente por medio del estadístico Eta al cuadrado parcial y «Pruebas de homogeneidad» para ver la homogeneidad de varianzas.

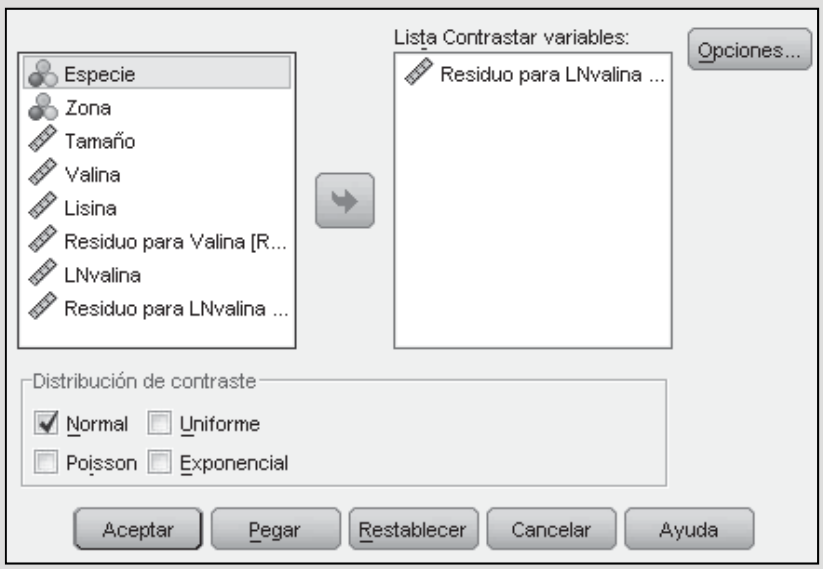


Paso 29. Después de pulsar «Aceptar» en la ventana del paso 24, se obtienen los resultados. En primer lugar hay que comprobar el requisito de que los residuos tengan una distribución Normal. Para ello se selecciona «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «K-S de 1 muestra…».



CUADRO IX.3. *(Continuación)*[SPSS]

Paso 30. En «Lista Contrastar variables:» se introducen los residuos y en «Distribución de contraste» se marca «Normal».



Paso 31. El test de Kolmogorov-Smirnov muestra que los residuos tienen una distribución Normal (p = 0,944). Si los residuos no tuviesen una distribución Normal, sería necesario elegir otra transformación como se explicó en los pasos 16 y 17.

Prueba de Kolmogorov-Smirnov para una muestra

		Residuo para LnValina
N		238
Parámetros normales[a,,b]	Media	,0000
	Desviación típica	,73201
Diferencias más extremas	Absoluta	,034
	Positiva	,025
	Negativa	-,034
Z de Kolmogorov-Smirnov		,527
Sig. asintót. (bilateral)		,944

a. La distribución de contraste es la Normal.

Paso 32. Los resultados del test de Levene muestran que también existe homogeneidad de varianzas (p = 0,092). Por lo tanto, se cumplen los dos requisitos necesarios para dar validez a los resultados obtenidos en el ANOVA.

CUADRO IX.3. *(Continuación)*[SPSS]

Contraste de Levene sobre la igualdad de las varianzas error[a]

Variable dependiente:LnValina

F	gl1	gl2	Sig.
1,290	5	232	,269

Paso 33. Los resultados del ANOVA se muestran en la siguiente tabla. Siguen sin existir diferencias significativas entre especies (ANOVA, $F_{1,\ 232}$ = 0,94, p = 0,333), pero existen diferencias significativas entre zonas (ANOVA, $F_{2,\ 232}$ = 8,9, p < 0,001). La interacción de ambos factores no es significativa (ANOVA, $F_{2,\ 232}$ = 1,9, p = 0,15).

Pruebas de los efectos inter-sujetos

Variable dependiente:LnValina

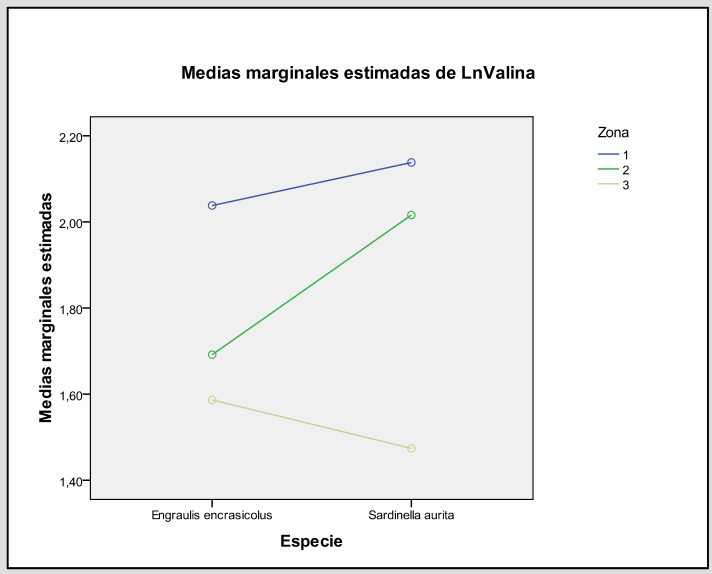
Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial
Modelo corregido	13,525[a]	5	2,705	4,942	,000	,096
Intersección	634,847	1	634,847	1159,766	,000	,833
Especie	,515	1	,515	,940	,333	,004
Zona	9,743	2	4,871	8,899	,000	,071
Especie * Zona	2,091	2	1,046	1,910	,150	,016
Error	126,995	232	,547			
Total	898,032	238				
Total corregida	140,520	237				

a. R cuadrado = ,096 (R cuadrado corregida = ,077)

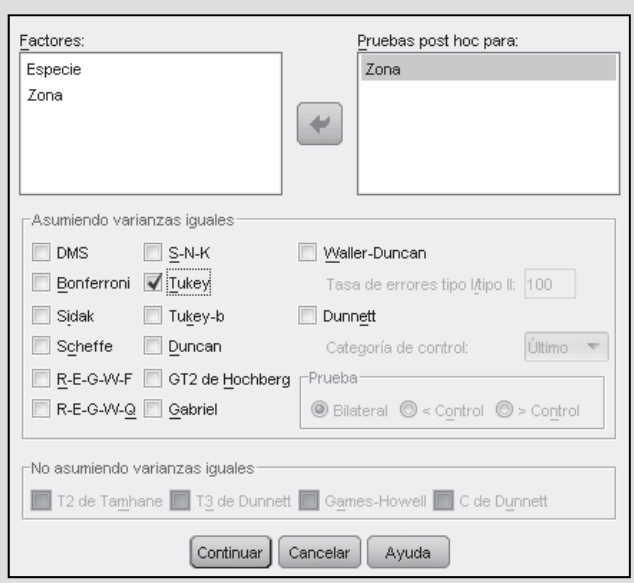
El estadístico Eta cuadrado parcial informa sobre el porcentaje de la variabilidad residual de la variable dependiente (no explicada ya por otras variables) que es explicada por cada factor, en un rango de 0 a 1. Aunque existen diferencias significativas en la concentración de Valina entre zonas, este factor no tiene una gran influencia en la concentración de Valina por larva, ya que el valor de Eta cuadrado parcial es muy bajo (0,071 para el factor «Zona»).

Paso 34. El gráfico de medias marginales permite visualizar la interacción significativa que existe entre los factores. Las líneas son relativamente paralelas entre la zona 1 y la 2, indicando poca interacción. Por el contrario, las diferencias de concentración de Valina por larva entre las dos especies son grandes entre la zona 2 y la 3. Para *E. encrasicolus* hay muy poca diferencia, pero la diferencia es importante en *S. aurita*, lo cual indica una cierta interacción entre los factores «Especie» y «Zona». En todo caso la interacción no debe considerarse ya que el factor «Especie» no es significativo.

CUADRO IX.3. *(Continuación)*[SPSS]



Paso 35. Como el factor «Zona» tiene más de 2 niveles (hay tres zonas) y es significativo, es posible hacer pruebas *post hoc* para ver cuáles son las zonas distintas. Para ello en la ventana del paso 18 seleccionamos «Post hoc…» y aparece la siguiente ventana. En «Pruebas post hoc para:» se introduce el factor «Zona». Existen pruebas para cuando las varianzas son iguales o son diferentes. Para varianzas iguales, que es nuestro caso, uno que se utiliza frecuentemente es el test «Tukey». Para el caso en que las varianzas no son homogéneas, el que más se usa es el test «T2 de Tamhane». Cualquiera de las pruebas es en general igualmente válida, excepto DMS, que no está corregida para evitar el efecto de las comparaciones múltiples.



CUADRO IX.3. *(Continuación)*[SPSS]

Paso 36. Las pruebas *post hoc* del test de Tukey muestran diferencias significativas en la cantidad del aminoácido Valina (en pg larva^{-1}) entre la zona 1 y la 3 (p < 0,001), entre la 2 y la 3 (p = 0,013), pero no hay diferencias entre la 1 y la 2 (p = 0,188). Por lo tanto, las diferencias observadas entre zonas se deben a que la zona 3 es muy diferente de las otras dos zonas.

Comparaciones múltiples

LnValina
DHS de Tukey

(I)Zona	(J)Zona	Diferencia de medias (I-J)	Error típ.	Sig.	Intervalo de confianza 95%	
					Límite inferior	Límite superior
1	2	,2144	,12239	,188	-,0743	,5031
	3	,5332*	,12297	,000	,2431	,8232
2	1	-,2144	,12239	,188	-,5031	,0743
	3	,3188*	,11155	,013	,0557	,5819
3	1	-,5332*	,12297	,000	-,8232	-,2431
	2	-,3188*	,11155	,013	-,5819	-,0557

Basadas en las medias observadas.
El término de error es la media cuadrática(Error) = ,547.

*. La diferencia de medias es significativa al nivel ,05.

IX.1.5. ANÁLISIS DE COVARIANZA

El análisis de la covarianza (ANCOVA) es una técnica estadística cuyo objetivo es eliminar el efecto de posibles variables que puedan estar relacionadas con la variable dependiente. Para controlar estas variables se realiza un análisis de varianza, en el que la variable dependiente es el error en los pronósticos- o residuo- al realizar un análisis de regresión lineal con las covariables como variables independientes y la variable dependiente será igual a la de nuestro ANOVA inicial.

En el ANCOVA la interpretación de sus resultados, al igual que en el ANOVA, se basa en los efectos de los factores sobre nuestras variables dependientes y de las interacciones entre los factores estudiados. Según Steel & Torrie (1985) este análisis se puede usar para:

1. Controlar errores y aumentar la precisión de nuestros análisis.
2. Ajustar las medias de las variables dependientes a las diferencias con los valores de las variables independientes usadas.
3. Ayudar en la interpretación de nuestros datos, especialmente en lo referente a la naturaleza de los efectos de los tratamientos.
4. Particionar una covarianza total o suma de productos cruzados en componentes.

5. Estimar datos que faltan.

Para poder interpretar de forma correcta un ANCOVA es necesario primero realizar el ANOVA, es decir, realizar el análisis sólo teniendo en cuenta la variable dependiente y el o los factores, sin considerar la(s) covariable(s). Esto se debe a que la interpretación del ANCOVA difiere en función de si los resultados obtenidos en el ANCOVA son iguales o diferentes a los obtenidos en el ANOVA.

En el ANCOVA toda variable que no tenga un efecto significativo puede ser eliminada del análisis. Si todas las covariables utilizadas poseen efectos no significativos, las conclusiones del ANCOVA y del ANOVA deben ser muy similares.

Si una o varias covariables tienen efectos significativos pueden ocurrir dos situaciones:

1. Que el resultado del ANOVA y del ANCOVA sea el mismo. Eso significa que el efecto de los factores sobre la variable dependiente no se modifica y, por tanto, el posible efecto de la(s) covariable(s) sobre la variable dependiente no afecta a la relación de esta última con los factores.
2. Si el resultado del ANOVA y del ANCOVA es distinto, puede deberse a dos motivos:
   2a. Puede ser debido a que un efecto significativo del ANOVA pasa a no ser significativo y, entonces, el efecto detectado en el ANOVA se debe al efecto de la(s) covariable(s) y no a la(s) variable(s) independientes (los factores).
   2b. También puede ser debido a que un efecto no significativo del ANOVA pase a significativo, lo cual indica que el factor, aun no estando relacionado con la variable dependiente globalmente considerada, sí correlaciona con la parte de la variable dependiente no explicada o no atribuible a la(s) covariable(s).

En el Cuadro IX.4 se explica como realizar un análisis de covarianza (ANCOVA).

CUADRO IX.4. Análisis de covarianza (ANCOVA)

Aplicación con R.

EJEMPLO. En el ejemplo usado en el análisis de varianza (Cuadro IX.3), se observó que había diferencias significativas en la longitud de la aleta anal entre las dos especies de tiburones (*Scyliorhinus canicula* y *Galeus melastomus*), pero no había diferencias entre zonas e investigadores. Sin embargo, es posible que la diferencias observadas se deban en realidad a diferentes tamaños de los individuos y no a que la aleta anal sea diferente. Para descartar el posible efecto del tamaño y determinar si efectivamente hay diferencias significativas en la longitud de la aleta anal entre especies, y también corroborar que no hay diferencias entre investigadores y/o zonas, es necesario realizar un ANCOVA considerando la longitud total como covariable. Los datos están en el archivo **Cuadro IX.3.V.csv**.

CUADRO IX.4. *(Continuación)*[R]

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro IX.4.R**, que se muestra en la siguiente ventana. Sólo es necesario especificar lo que se solicita en la sección modificable por el usuario. Para que este *script* funcione correctamente es necesario tener instalados los paquetes "nortest", "mvtnorm", "class", "VGAM", "splines", "stats4", "lawstat", "car" "multcomp", "survival" y "e1071", o realizar la instalación como se explica en el Apéndice II.

```
#Cuadro IX.4
#Análisis de covarianza.

################################################
#Sección modificable por el usuario
################################################

#Lectura de datos.
datos<-read.csv2("Cuadro IX.3.V.csv",header=TRUE,encoding="latin1")

#Selección variable dependiente. En caso de que se deseen utilizar val
varInteres<-c("Longitud.anal")
#varInteres<-c("Diámetro.ojo")

#Selección de un tipo de transformación para las variables de interés.
varInteres1X2<-NULL
varInteres1X<-NULL
varInteresLN<-NULL
varInteresLOG10<-NULL
varInteresSQR<-NULL
varInteresX2<-NULL
varInteresX3<-NULL
varInteresEX<-NULL
varInteresASN<-NULL

#varInteres1X2<-c(1,1)
#varInteres1X<-c(1,1)
#varInteresLN<-c(1,1)
#varInteresLOG10<-c(1,1)
#varInteresSQR<-c(1,1)
#varInteresX2<-c(1,1)
#varInteresX3<-c(1,1)
#varInteresEX<-c(1,1)
#varInteresASN<-c(1,1)
```

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro IX.3.V.csv**.

```
datos<-read.csv2("Cuadro IX.3.V.csv",header=TRUE,encoding="latin1")
```

Paso 2. Se selecciona la variable dependiente, en este caso la longitud de la aleta anal.

```
varInteres<-c("Longitud.anal")
#varInteres<-c("Diámetro.ojo")
```

CUADRO IX.4. *(Continuación)*[R]

Paso 3. En caso de que sea necesario es posible transformar la variable dependiente; para ello simplemente hay que quitar el símbolo # en el tipo de transformación que se quiera realizar. En primer lugar se hace el análisis sin transformar los datos, y en el caso de no cumplirse la Normalidad y/o homogeneidad de varianzas, entonces se realizaría el análisis aplicando la transformación más adecuada.

```
#varInteres1X2<-c(1,1)
#varInteres1X<-c(1,1)
#varInteresLN<-c(1,1)
#varInteresLOG10<-c(1,1)
#varInteresSQR<-c(1,1)
#varInteresX2<-c(1,1)
#varInteresX3<-c(1,1)
#varInteresEX<-c(1,1)
#varInteresASN<-c(1,1)
```

Paso 4. A continuación se indica cuales son los factores con los que se va a trabajar, hasta un máximo de 3 factores con este *script*. En este caso se hace el análisis con los tres factores: «Especie», «Investigador» y «Zona». Si no se quiere considerar algún factor, simplemente se activa la opción «NULL» quitando el símbolo # delante de la instrucción.

```
Factor1<-c("Especie")
Factor2<-c("Investigador")
#Factor2<-NULL
Factor3<-c("Zona")
#Factor3<-NULL
```

Paso 5. Es posible trabajar con 4 covariables, aunque en el ejemplo sólo se considera la «Longitud.total».

```
Covariable1<-c("Longitud.total")
Covariable2<-c("Diámetro.ojo")
Covariable2<-NULL
Covariable3<-c("Base.anal")
Covariable3<-NULL
Covariable4<-c("")
Covariable4<-NULL
```

Paso 6. Un requisito importante en el análisis de covarianza es que la relación entre la variable dependiente, en este caso la longitud de la aleta anal, con la covariable o covariables sea lineal. Para comprobar si la relación es lineal, es posible aplicar una regresión entre la variable dependiente y las covariables, y también representar gráficamente los datos. En este caso sólo se considera la «Longitud.total».

CUADRO IX.4. *(Continuación)*[R]

```
Covregr1<-c("Longitud.total")
#Covregr1<-NULL
Covregr2<-c("Diámetro.ojo")
Covregr2<-NULL
Covregr3<-c("Base.anal")
Covregr3<-NULL
Covregr4<-c("")
Covregr4<-NULL
```

Si la relación no fuera lineal, es posible transformar logarítmicamente las covariables. En el ejemplo, como se deja activada «transcov<-c(0)», no se realiza ninguna transformación.

```
#1.Transformar covariable 1
#2.Transformar covariable 2
#3.Transformar covariable 3
#4.Transformar covariable 4

transcov<-c(1,2)
transcov<-c(0)
```

Paso 7. También es posible seleccionar dos grupos en dos variables diferentes. En este ejemplo no se hace ninguna selección.

```
#Variables de selección.
#varSeleccion1<-list(c("",""))
varSeleccion1<-NULL
#varSeleccion2<-list(c("","",""))
varSeleccion2<-NULL
```

Paso 8. Es posible realizar el análisis con el tipo II o el III. Realizaremos el análisis con el tipo III de suma de cuadrados.

```
#SS<-c(2)
SS<-c(3)
```

Paso 9. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
ArchivodeSalida<-"Salida Cuadro IX.4.txt"
```

Paso 10. En primer lugar nos fijamos en la Normalidad de los residuos y la homogeneidad de varianzas. Se observa que, tanto el test Kolmogorov–Smirnov con la corrección de Lilliefors (p = 0,256) como el de Shapiro-Wilk (p = 0,34) presentan valores de $p > 0,05$ y, por tanto, se acepta la hipótesis nula de bondad de ajuste y los residuos se ajustan a una distribución Normal. El test de Levene considerando la mediana muestra que no se cumple la homogeneidad de varianzas para el factor 1 (p = 0,002), pero si para el factor 2 (p = 0,725) y el factor 3 (p = 0,623). Por lo tanto, es necesario transformar los datos.

CUADRO IX.4. *(Continuación)*[R]

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.0921, p-value = 0.2556


[[7]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9771, p-value = 0.3400


[[8]]
[[8]][[1]]
[1] "Factor1"

[[8]][[2]]
Levene's Test for Homogeneity of Variance
      Df F value   Pr(>F)
group  1  10.227 0.002278 **
      56
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[[8]][[3]]
[1] "Factor2"

[[8]][[4]]
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  1  0.1246 0.7254
      56

[[8]][[5]]
[1] "Factor3"

[[8]][[6]]
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  3  0.5923 0.6227
      54
```

Paso 11. Nos fijamos en la asimetría, la cual es negativa (-0,328), pero la curtosis es positiva (0,698). Por tanto, en base a la Figura IX.1, se podría escoger la transformación «raíz» o «X^2». Se repite el análisis pero transformando los datos y, para ello, en el paso 3 quitaríamos el símbolo # de la transformación que queremos realizar. En este caso probamos haciendo la transformación de la raíz cuadrada.

```
[[9]]
[1] "Sesgo"                "-0.328227935965377"

[[10]]
[1] "Curtosis"            "0.698715658423597"
```

CUADRO IX.4. *(Continuación)*[R]

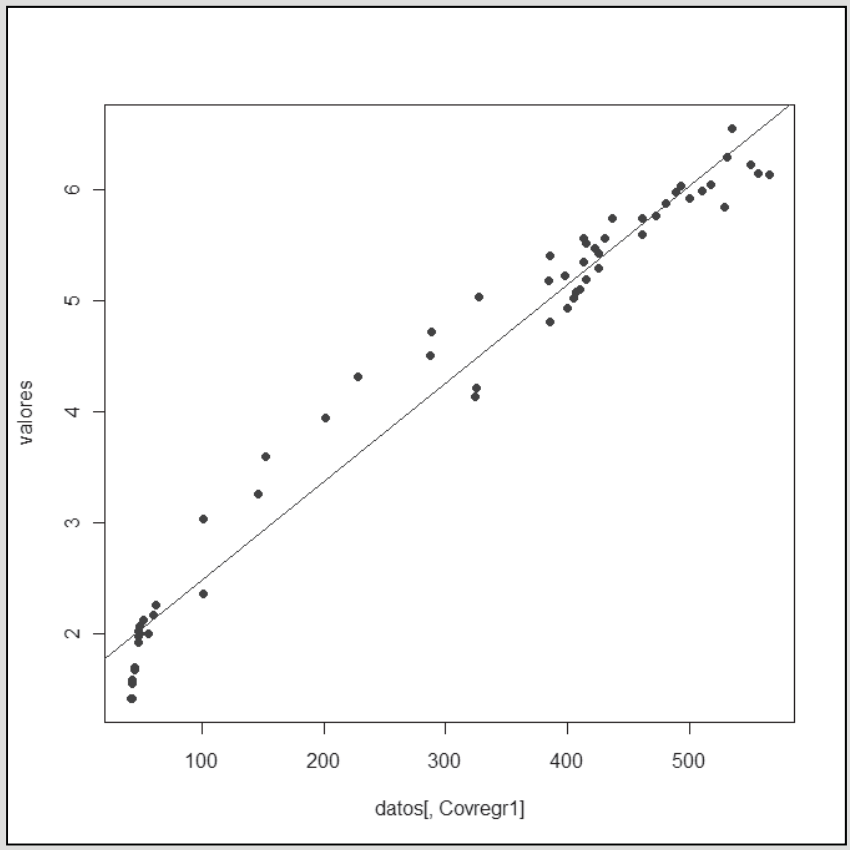
Paso 12. Ahora se observa que, tanto el test Kolmogorov–Smirnov con la corrección de Lilliefors (p = 0,576) como el de Shapiro-Wilk (p = 0,361) tienen valores de p > 0,05 y, por tanto, se acepta la hipótesis nula de bondad de ajuste y los residuos se ajustan a una distribución Normal. El test de Levene considerando la mediana muestra que se cumple la homogeneidad de varianzas para el factor 1 (p = 0,263), para el factor 2 (p = 0,202) y para el factor 3 (p = 0,159).

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.075, p-value = 0.5761


[[7]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9777, p-value = 0.3613


[[8]]
[[8]][[1]]
[1] "Factor1"

[[8]][[2]]
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  1  1.2781 0.2631
      56

[[8]][[3]]
[1] "Factor2"

[[8]][[4]]
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  1  1.6696 0.2016
      56

[[8]][[5]]
[1] "Factor3"

[[8]][[6]]
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group  3  1.7961 0.1589
      54
```

Paso 13. La regresión entre la covariable y la variable dependiente se ajusta bien a una relación lineal ya que r^2 = 0,97 y en la siguiente página se muestra el gráfico entre ambas variables.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.6004222  0.0775458   20.64   <2e-16 ***
datos[, Covregr1] 0.0088698  0.0002162   41.02   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3039 on 56 degrees of freedom
Multiple R-squared: 0.9678,     Adjusted R-squared: 0.9672
F-statistic:  1683 on 1 and 56 DF,  p-value: < 2.2e-16
```

CUADRO IX.4. *(Continuación)*[R]



Paso 14. La tabla de ANOVA muestra que existe una clara relación entre la longitud total y la longitud de la aleta anal (ANCOVA, $F_{1, 41}$ = 576,37, p < 0,001). Al considerar el efecto de la longitud total siguen existiendo claras diferencias en la longitud de la aleta anal entre especies (ANOVA, $F_{1, 41}$ = 24,09, p < 0,001), pero no entre investigadores (ANOVA, $F_{1, 41}$ = 1,44, p = 0,235) ni entre zonas (ANOVA, $F_{3, 41}$ = 1,06, p = 0,374).

```
[[4]]
Anova Table (Type III tests)

Response: datos2$valores
                            Sum Sq Df  F value    Pr(>F)
(Intercept)                  1.987  1  32.4740 1.160e-06 ***
datos2$C1                   35.265  1 576.3747 < 2.2e-16 ***
datos2$F1                    1.474  1  24.0953 1.504e-05 ***
datos2$F2                    0.089  1   1.4474 0.2358415
datos2$F3                    0.196  3   1.0652 0.3743666
datos2$F1:datos2$F2          0.574  1   9.3866 0.0038534 **
datos2$F1:datos2$F3          1.483  3   8.0818 0.0002405 ***
datos2$F2:datos2$F3          0.301  3   1.6415 0.1946235
datos2$F1:datos2$F2:datos2$F3 1.023  3   5.5707 0.0026676 **
Residuals                    2.509 41
```

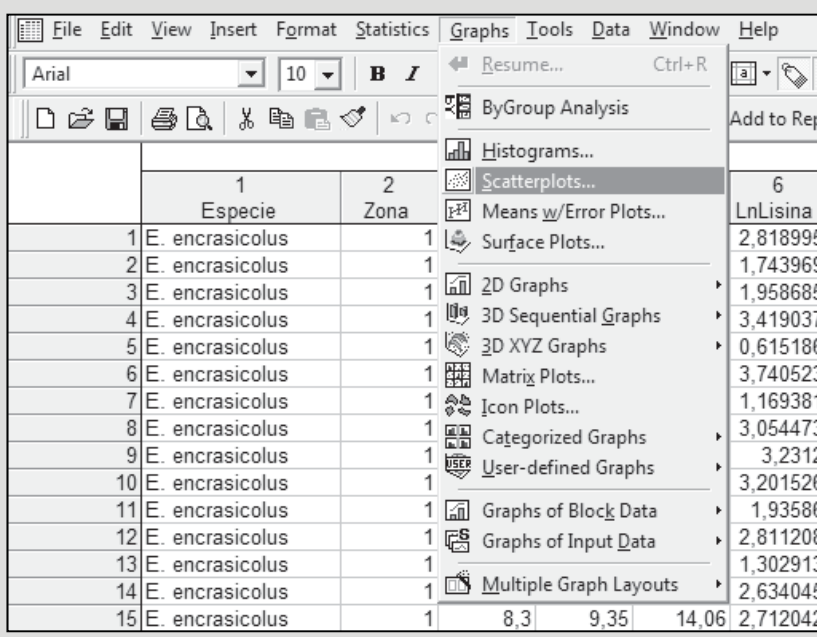
CUADRO IX.4. *(Continuación)*[STATISTICA]

Aplicación con STATISTICA.

EJEMPLO. En el caso utilizado en el análisis de varianza (Cuadro IX.3), se observó que había diferencias significativas en la concentración del aminoácido Lisina (en pg larva^{-1}), en diferentes zonas del Canal de Sicilia. Sin embargo, estas diferencias podrían ser debidas al diferente tamaño de las larvas muestreadas, y no necesariamente al contenido de aminoácidos por larva entre especies y/o zonas. Para descartar el posible efecto del tamaño y determinar si efectivamente hay diferencias significativas en el contenido de aminoácidos entre las especies y/o las zonas, es necesario realizar un ANCOVA considerando al tamaño como covariable. Los datos están en el archivo **Cuadro IX.4.sta**.

Paso 1. Para poder interpretar mejor un ANCOVA es necesario primero hacer un ANOVA. Por tanto, se tienen que seguir en primer lugar todos los pasos descritos en el Cuadro IX.3 en la aplicación con STATISTICA. En el Cuadro IX.3 se observó que era necesario aplicar el logaritmo a la variable «Lisina» para cumplir con los requisitos de Normalidad y homogeneidad de varianzas. Por tanto, en el análisis de ANCOVA usaremos la variable transformada «LnLisina».

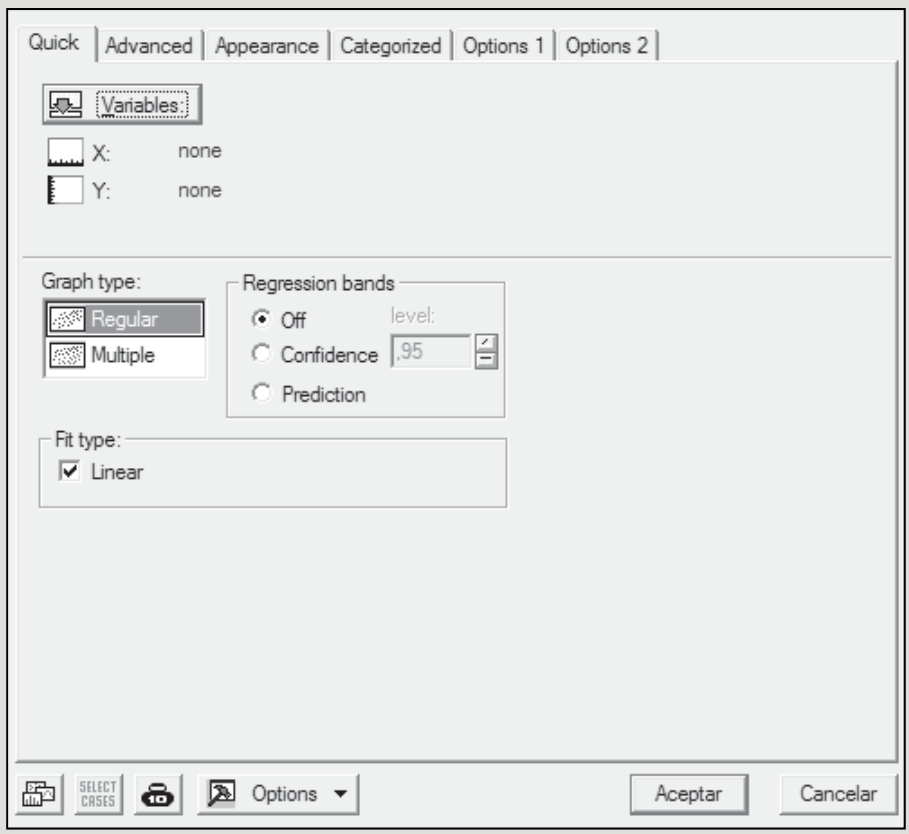
Paso 2. Debido a que el ANCOVA realiza un análisis de regresión lineal, debemos comprobar que la relación entre nuestras variables sea lineal; de no ser así debemos proceder a transformar los datos antes de realizar cualquier análisis. En primer lugar es importante graficar los resultados para ver el tipo de relación que existe entre las variables. Para ello vamos a «Graphs» ⇨ «Scatterplots…».

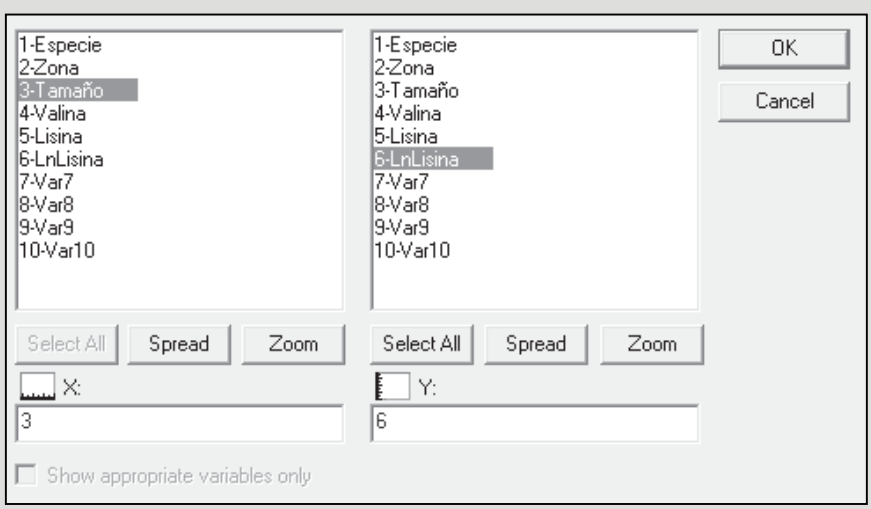


	1 Especie	2 Zona				6 LnLisina
1	E. encrasicolus	1				2,818995
2	E. encrasicolus	1				1,743969
3	E. encrasicolus	1				1,958685
4	E. encrasicolus	1				3,419037
5	E. encrasicolus	1				0,615186
6	E. encrasicolus	1				3,740523
7	E. encrasicolus	1				1,169381
8	E. encrasicolus	1				3,054473
9	E. encrasicolus	1				3,2312
10	E. encrasicolus	1				3,201526
11	E. encrasicolus	1				1,93586
12	E. encrasicolus	1				2,811208
13	E. encrasicolus	1				1,302913
14	E. encrasicolus	1				2,634045
15	E. encrasicolus	1	8,3	9,35	14,06	2,712042

CUADRO IX.4. *(Continuación)*[STATISTICA]

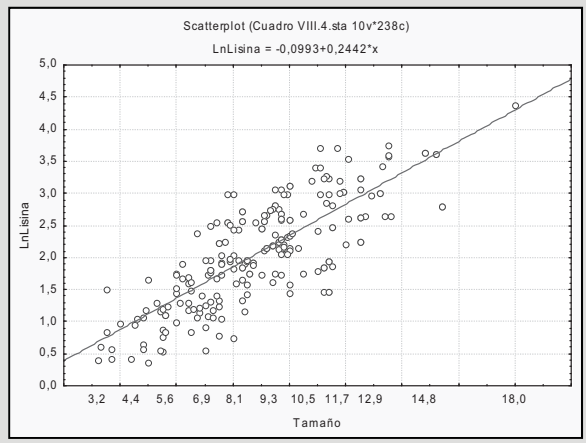
Paso 3. Al pulsar en «Variables» aparece la ventana que se muestra posteriormente donde seleccionamos las variables «LnLisina» y «Tamaño».





CUADRO IX.4. *(Continuación)*[STATISTICA]

Paso 4. La gráfica que se obtiene parece mostrar una relación lineal, aunque es conveniente realizar un análisis para comprobarlo.

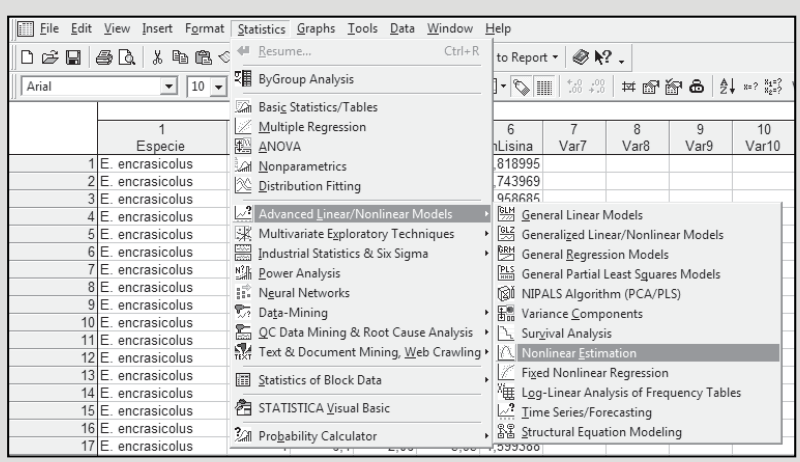


Paso 5. Para comprobar cual es el mejor modelo (Lineal, Logarítmico, Potencial o Exponencial) que relaciona el tamaño de las larvas con el LnLisina, hay que comparar los modelos y en el caso de que el ajuste fuese mejor con un modelo que no sea el Lineal, entonces se tendría que realizar la transformación como se muestra en la Tabla IX.1

Tabla IX.1. Transformaciones para conseguir relación lineal.

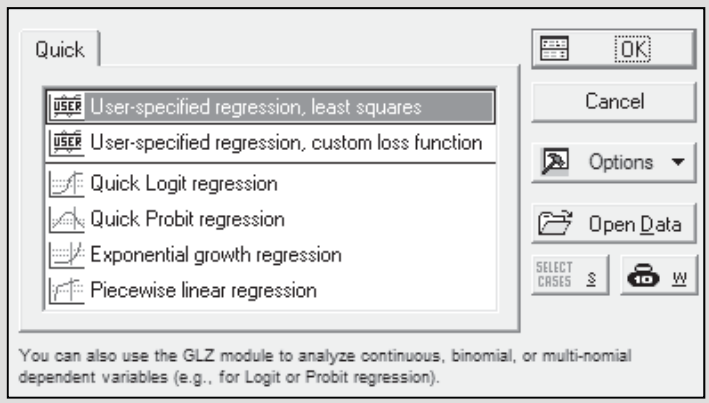
	Modelo	Transformación x	*Transformación* y
Logarítmico	$y = a + b\ln x$	$t(x) = \ln(x)$	$t(y) = y$
Potencial	$y = ax^b$	$t(x) = \ln(x)$	$t(y) = \ln(y)$
Exponencial	$y = ae^{bx}$	$t(x) = x$	$t(y) = \ln(y)$

Paso 6. Para ello se va a «Advanced Linear/Nonlinear Models» ⇨ «Nonlinear Estimation».

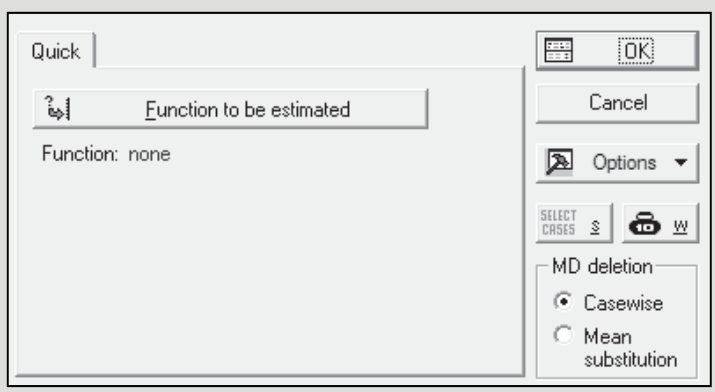


CUADRO IX.4. *(Continuación)*[STATISTICA]

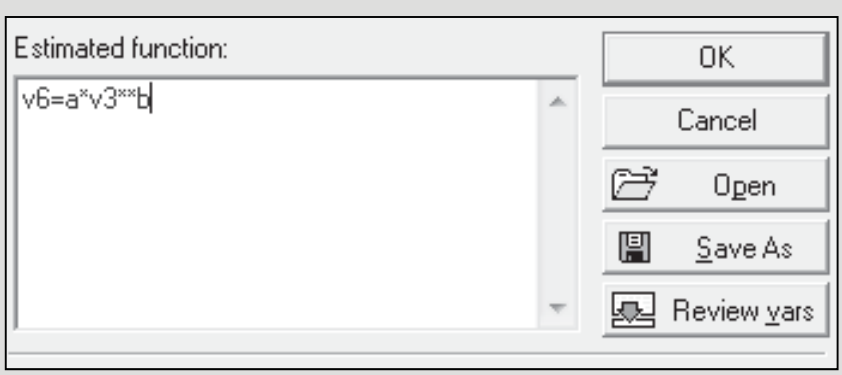
Paso 7. En la ventana siguiente se selecciona «User-specified regression, least squares».



Paso 8. A continuación se selecciona «Function to be estimated».



Paso 9. Ponemos la ecuación del modelo que queremos estimar. En la ventana se muestra la ecuación para el modelo Potencial «v6=a*v3**b». Para el modelo Exponencial sería «v6=a*exp(b*v3)», para el modelo Lineal «v6=a +b*v3» y para el modelo Logarítmico «v6=a+b*log(v3)».



CUADRO IX.4. *(Continuación)*[STATISTICA]

Paso 10. Después de pulsar tres veces «OK» aparecen los resultados. Hay que repetir el proceso para cada modelo. Se observa que el valor de *r* más alto es para el modelo Lineal (0,794) y, por tanto, no es necesario realizar ningún tipo de transformación. En cualquier caso, aunque el valor de *r* del modelo lineal no fuera el más alto, si las diferencias son tan pequeñas como en este ejemplo, tampoco justificaría realizar ningún tipo de transformación. Si fuese necesario transformar las variables, ello se realizaría siguiendo los pasos 10 y 11 del Cuadro IX.1 en la aplicación con STATISTICA.

Modelo Lineal

```
Model is: v6=a+b*v3
Dependent variable: LnLisina          Independent variables: 1
Loss function: least squares
Final value: 55,02636955
Proportion of variance accounted for: ,63087598    R =,79427702
```

Modelo Logarítmico

```
Model is: v6=a+b*log(v3)
Dependent variable: LnLisina          Independent variables: 1
Loss function: least squares
Final value: 55,3044631
Proportion of variance accounted for: ,62901049    R =,79310182
```

Modelo Potencial

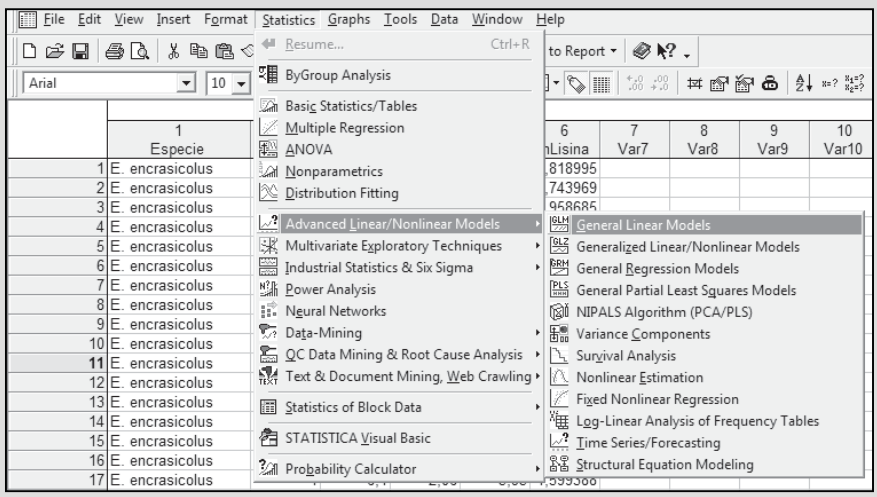
```
Model is: v6=a*v3**b
Dependent variable: LnLisina          Independent variables: 1
Loss function: least squares
Final value: 55,17513846
Proportion of variance accounted for: ,62987802    R =,79364855
```

Modelo Exponencial

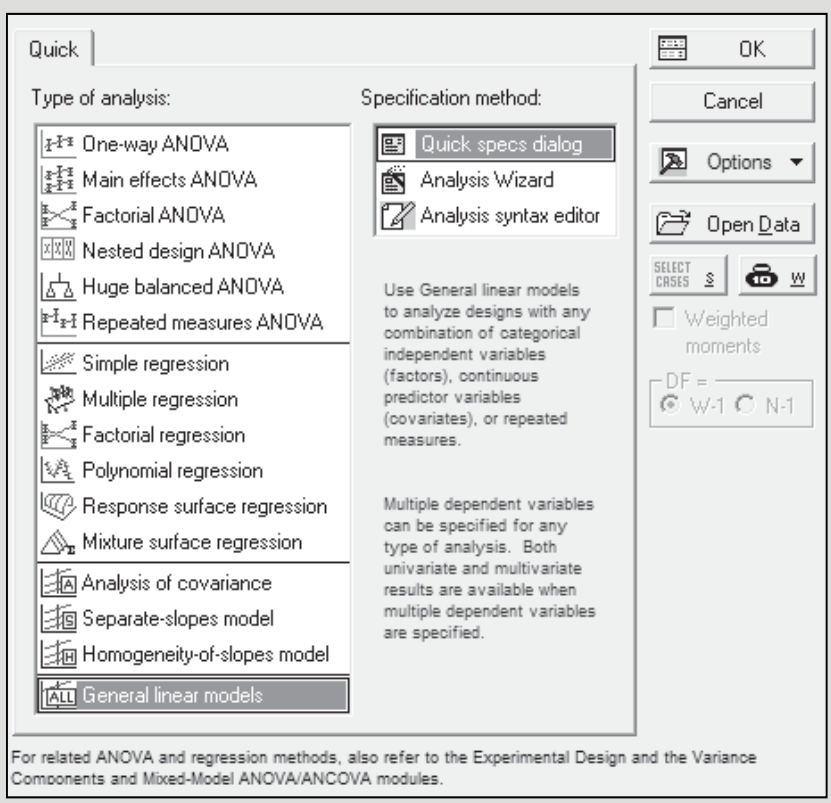
```
Model is: v6=a*exp(b*v3)
Dependent variable: LnLisina          Independent variables: 1
Loss function: least squares
Final value: 62,46480071
Proportion of variance accounted for: ,58097802    R =,76221914
```

CUADRO IX.4. *(Continuación)*[STATISTICA]

Paso 11. Para realizar el ANCOVA se selecciona «Statistics» ⇨ «Advanced Linear/Nonlinear Models» ⇨ «General Linear Models».

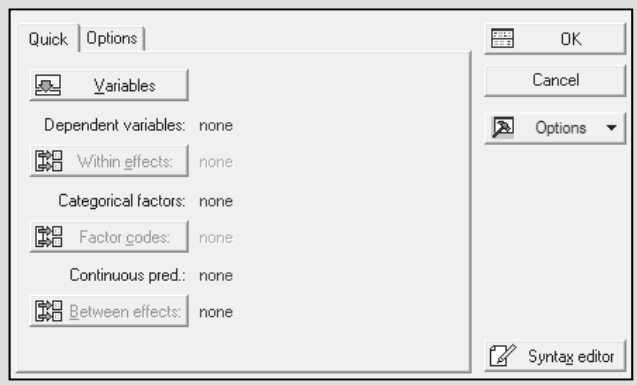


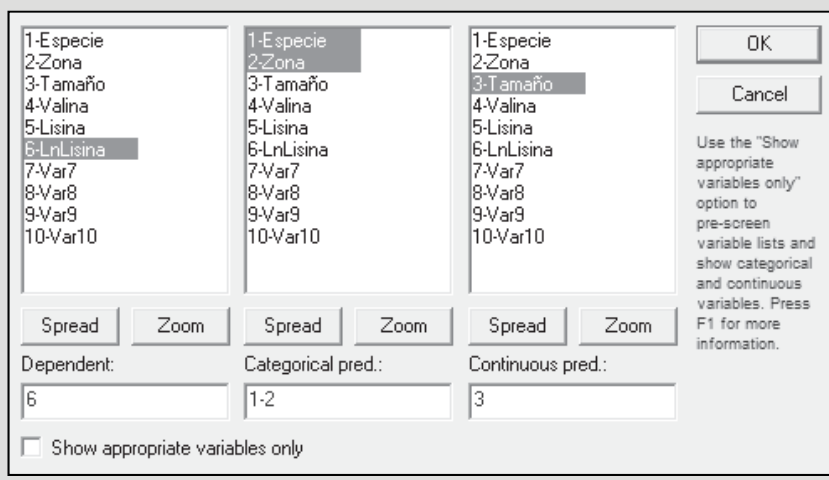
Paso 12. Posteriormente se selecciona «General linear models».

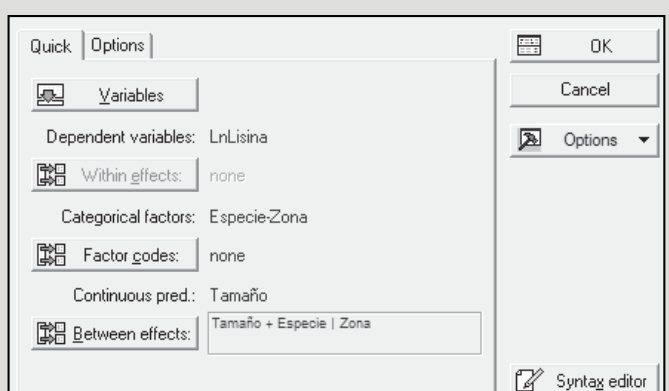

Quick
Type of analysis:
One-way ANOVA
Main effects ANOVA
Factorial ANOVA
Nested design ANOVA
Huge balanced ANOVA
Repeated measures ANOVA
Simple regression
Multiple regression
Factorial regression
Polynomial regression
Response surface regression
Mixture surface regression
Analysis of covariance
Separate-slopes model
Homogeneity-of-slopes model
General linear models
Specification method:
Quick specs dialog
Analysis Wizard
Analysis syntax editor
Use General linear models to analyze designs with any combination of categorical independent variables (factors), continuous predictor variables (covariates), or repeated measures.
Multiple dependent variables can be specified for any type of analysis. Both univariate and multivariate results are available when multiple dependent variables are specified.
OK
Cancel
Options
Open Data
Weighted moments
DF =
W-1
N-1
For related ANOVA and regression methods, also refer to the Experimental Design and the Variance Components and Mixed-Model ANOVA/ANCOVA modules.

CUADRO IX.4. *(Continuación)*[STATISTICA]

Paso 13. Al pulsar en «Variables» aparece la segunda ventana que se muestra a continuación donde seleccionamos la variable dependiente «LnLisina», los factores «Especie» y «Zona» y la covariable «Tamaño».

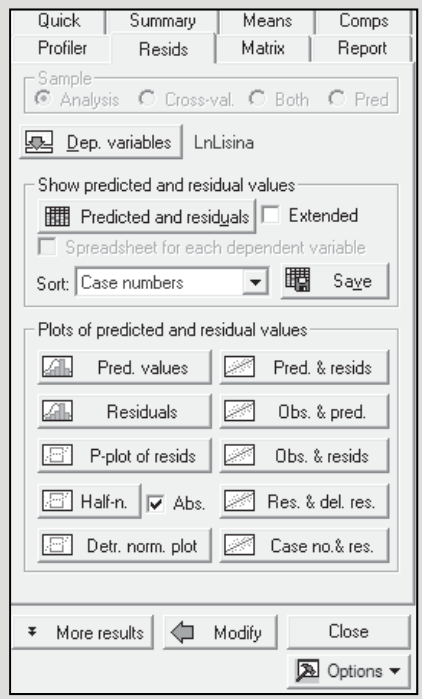




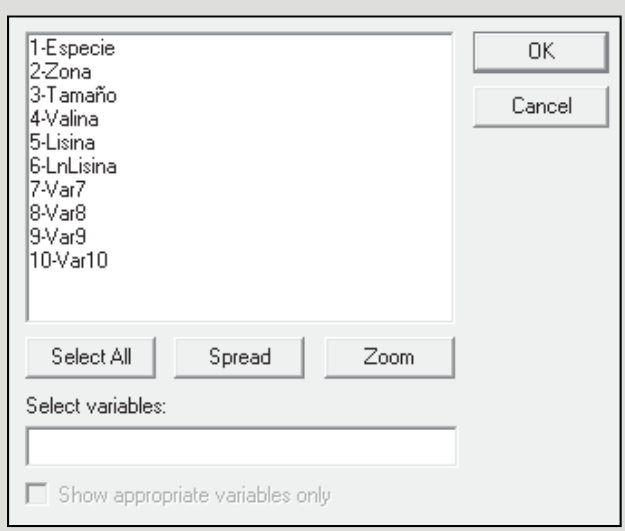


CUADRO IX.4. *(Continuación)*[STATISTICA]

Paso 14. Al pulsar en «OK» obtenemos la siguiente ventana. Para solventar cualquier duda con respecto al requisito de distribución Normal, es conveniente ver si los residuos tienen distribución Normal. En la pestaña «Resids» pulsamos en «Save» para guardar los residuos.



Paso 15. En la ventana que aparece pulsamos «OK» sin seleccionar ninguna variable.

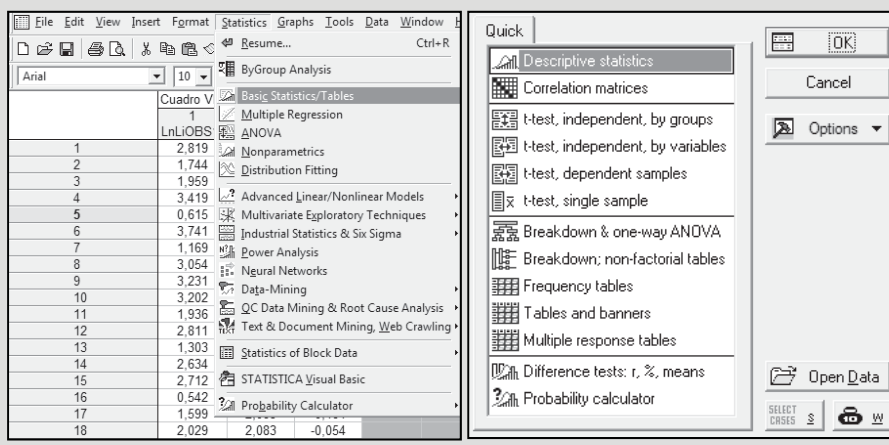


CUADRO IX.4. *(Continuación)*[STATISTICA]

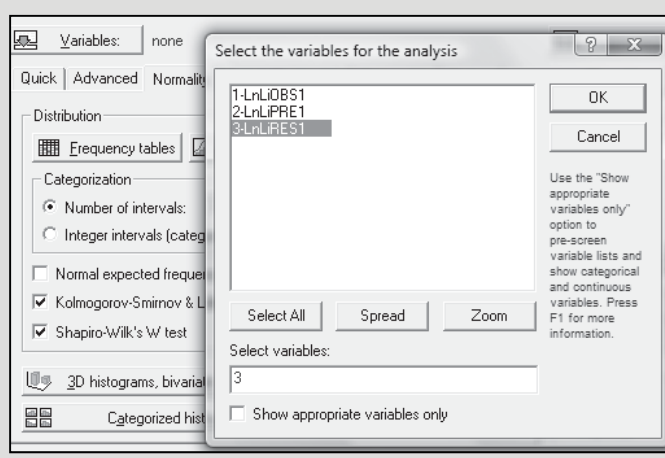
Paso 16. Se obtiene una tabla, en la cual la tercera columna son los residuos.

	Cuadro IX.4.sta		
	1 LnLiOBS1	2 LnLiPRE1	3 LnLiRES1
1	2,819	2,412	0,407
2	1,744	2,412	-0,668
3	1,959	1,866	0,092
4	3,419	3,312	0,107
5	0,615	0,942	-0,327
6	3,741	3,360	0,381
7	1,169	1,440	-0,270

Paso 17. Se selecciona «Statistics» ⇨ «Basic Statistics/Tables» ⇨ «Descriptive statistics».

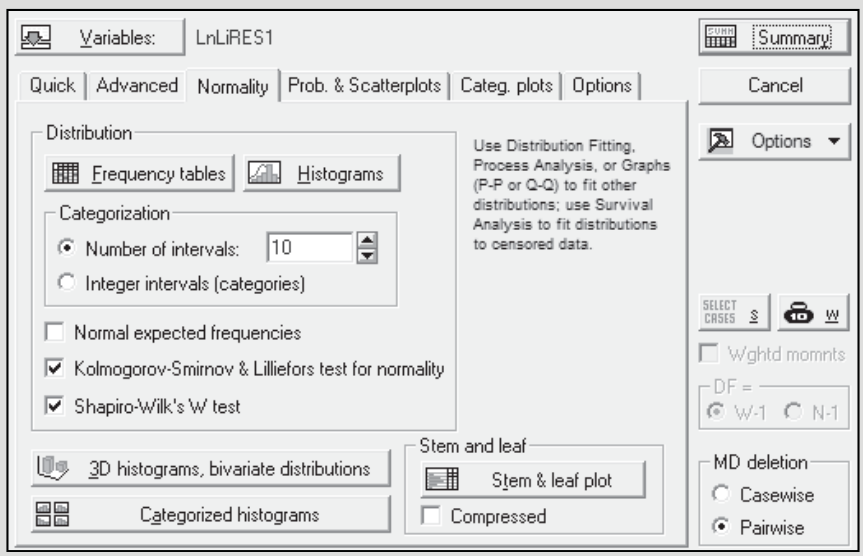


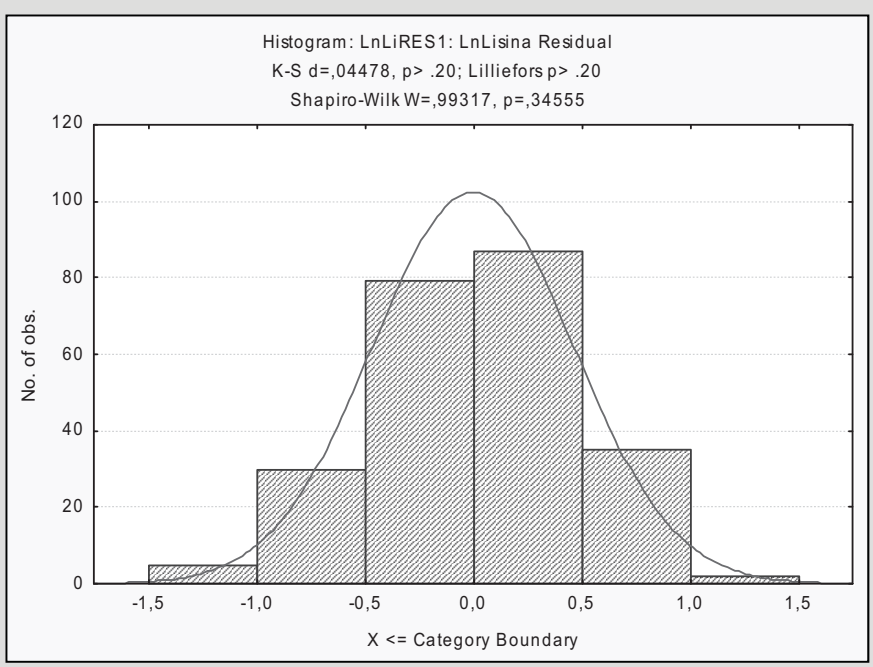
Paso 18. En la pestaña «Normality» marcamos los test «Kolmogorov-Smirnov & Lilliefors test for normality» y «Shapiro-Wilk´s test». En «Variables» se selecciona «LnLiRES1».



CUADRO IX.4. *(Continuación)*[STATISTICA]

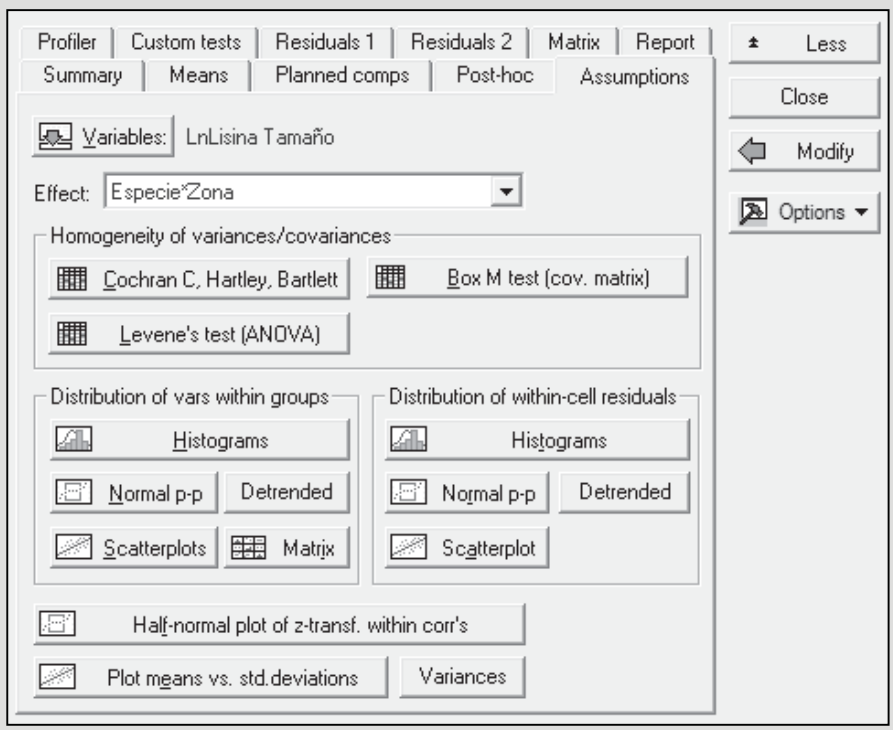
Paso 19. En la siguiente ventana al pulsar «Histograms» aparece el siguiente gráfico donde se observa que tanto el test Kolmogorov–Smirnov con la corrección de Lilliefors, como el de Shapiro-Wilk muestran valores de $p > 0{,}05$ y, por tanto, se acepta la hipótesis de que los residuos se ajustan a una distribución Normal.





CUADRO IX.4. *(Continuación)*[STATISTICA]

Paso 20. En la ventana del paso 14, pulsamos en «More results» y aparece la siguiente ventana donde, en la pestaña «Assumptions», pulsamos «Levene´s test (ANOVA)», para ver el requisito de homogeneidad de varianzas.

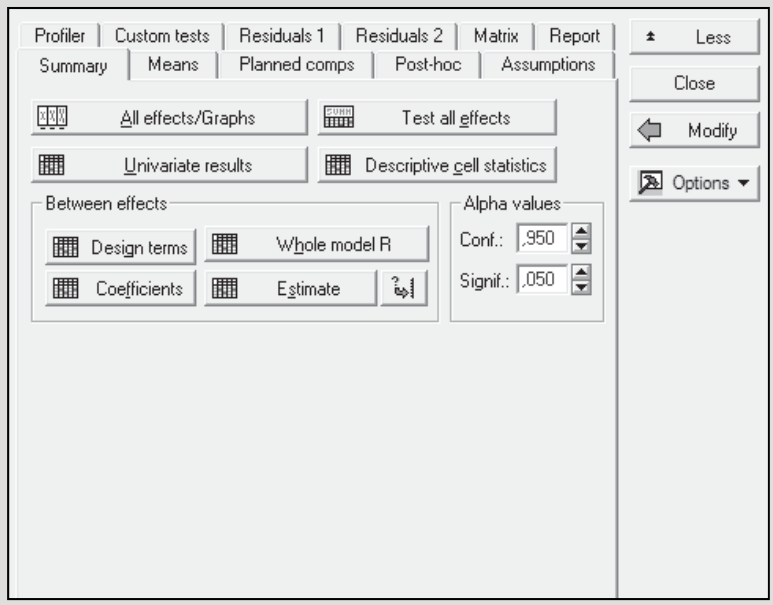


Paso 21. Aparece la tabla que se muestra a continuación. Como la probabilidad es mayor que 0,05, aceptamos la hipótesis nula de homogeneidad de varianzas, tanto entre factores (p = 0,393), como para la covariable (p = 0,293). Por tanto, se cumple el otro requisito del ANCOVA.

	Levene's Test for Homogeneity of Variances (Cuadro IX.4.sta) Effect: Especie*Zona Degrees of freedom for all F's: 5, 232			
	MS Effect	MS Error	F	p
LnLisina	0,221353	0,212179	1,043238	0,392945
Tamaño	2,984223	2,417515	1,234418	0,293716

Paso 22. En la ventana del paso 20, en la pestaña «Summary» pulsamos en «Univariate results» para obtener los resultados.

CUADRO IX.4. *(Continuación)* *STATISTICA*



Paso 23. Los resultados se muestran en la siguiente tabla. Existe una clara relación entre la concentración de Lisina (en pg larva^{-1}) y el tamaño (en mm) ($F_{1,\ 231}$ = 374,7, p < 0,001), por lo tanto el efecto de la covariable es significativo. También existen diferencias significativas en la concentración del aminoácido entre zonas ($F_{2,\ 231}$ = 3,3, p = 0,038).

	Univariate Results for Each DV (Cuadro IX.4.sta) Sigma-restricted parameterization Effective hypothesis decomposition				
Effect	Degr. of Freedom	LnLisina SS	LnLisina MS	LnLisina F	LnLisina p
Intercept	1	0,0373	0,03731	0,1692	0,681243
Tamaño	1	82,6535	82,65345	374,7492	0,000000
Especie	1	0,0719	0,07193	0,3262	0,568490
Zona	2	1,4616	0,73080	3,3134	0,038128
Especie*Zona	2	1,4702	0,73511	3,3330	0,037411
Error	231	50,9486	0,22056		
Total	237	149,0728			

Antes de considerar el efecto del tamaño, no existían diferencias significativas en la cantidad del aminoácido Lisina (en pg larva^{-1}), entre las especies *S. aurita* y *E. encrasicolus* (ANOVA, $F_{1,\ 232}$ = 1,53, p = 0,217) (véase paso 23 del Cuadro IX.3 en la aplicación con STATISTICA). Ahora, tampoco hay diferencias significativas (ANOVA, $F_{1,\ 231}$ = 0,32, p = 0,568). El efecto combinado de ambos factores («Especie» y «Zona») no lo podemos tener en cuenta, ya que el efecto individual del factor «Especie» sobre la variable dependiente «Lisina» no es significativo.

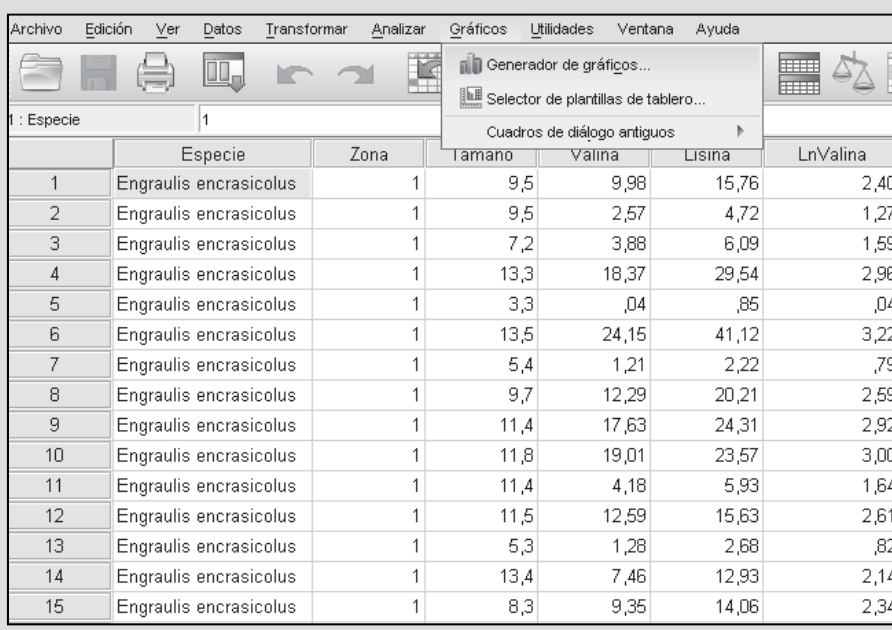
CUADRO IX.4. *(Continuación)*[SPSS]

Aplicación con SPSS.

EJEMPLO. En el ejemplo usado en el análisis de varianza (Cuadro IX.3), se observó que había diferencias significativas en la concentración del aminoácido Valina (en pg larva^{-1}), en diferentes zonas del Canal de Sicilia. Sin embargo, estas diferencias podrían ser debidas al diferente tamaño de las larvas muestreadas, y no necesariamente al contenido de aminoácidos por larva entre especies y/o zonas. Para descartar el posible efecto del tamaño y determinar si efectivamente hay diferencias significativas en el contenido de aminoácidos entre las especies y/o las zonas, es necesario realizar un ANCOVA considerando al tamaño como covariable. Los datos están en el archivo **Cuadro IX.4.sav**.

Paso 1. Para poder interpretar mejor un ANCOVA es necesario primero hacer un ANOVA. Por tanto, se tienen que seguir en primer lugar todos los pasos descritos en el Cuadro IX.3 en la aplicación con SPSS. En el Cuadro IX.3 se observó que era necesario aplicar el logaritmo a la variable «Valina» para cumplir con los requisitos de Normalidad y homogeneidad de varianzas. Por tanto, en el análisis de ANCOVA usaremos la variable transformada «LnValina».

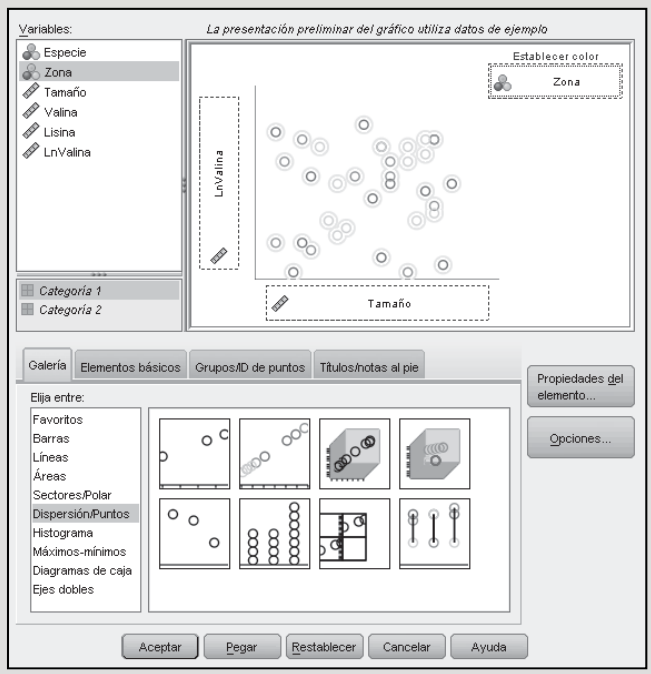
Paso 2. Debido a que el ANCOVA realiza un análisis de regresión lineal, debemos comprobar que la relación entre nuestras variables sea lineal, de no ser así debemos proceder a transformar los datos antes de realizar el análisis. En primer lugar es importante graficar los resultados para ver el tipo de relación que existe entre las variables. Para ello vamos a «Gráficos» ⇨ «Generador de gráficos».



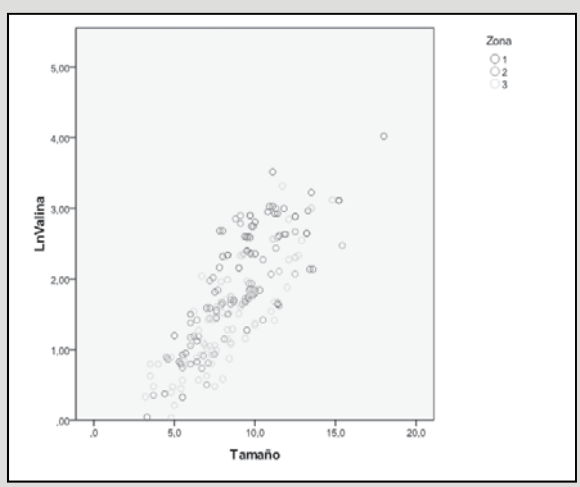
	Especie	Zona	Tamaño	Valina	Lisina	LnValina
1	Engraulis encrasicolus	1	9,5	9,98	15,76	2,40
2	Engraulis encrasicolus	1	9,5	2,57	4,72	1,27
3	Engraulis encrasicolus	1	7,2	3,88	6,09	1,59
4	Engraulis encrasicolus	1	13,3	18,37	29,54	2,96
5	Engraulis encrasicolus	1	3,3	,04	,85	,04
6	Engraulis encrasicolus	1	13,5	24,15	41,12	3,22
7	Engraulis encrasicolus	1	5,4	1,21	2,22	,79
8	Engraulis encrasicolus	1	9,7	12,29	20,21	2,59
9	Engraulis encrasicolus	1	11,4	17,63	24,31	2,92
10	Engraulis encrasicolus	1	11,8	19,01	23,57	3,00
11	Engraulis encrasicolus	1	11,4	4,18	5,93	1,64
12	Engraulis encrasicolus	1	11,5	12,59	15,63	2,61
13	Engraulis encrasicolus	1	5,3	1,28	2,68	,82
14	Engraulis encrasicolus	1	13,4	7,46	12,93	2,14
15	Engraulis encrasicolus	1	8,3	9,35	14,06	2,34

CUADRO IX.4. *(Continuación)*[SPSS]

Paso 3. En «Galería» se selecciona «Dispersión/Puntos» y entre los distintos gráficos que aparecen se escoge «Dispersión Agrupado». Posteriormente se arrastran las variables que queremos representar desde el cuadro «Variables» a los ejes X e Y de la gráfica y, por último, en este caso establecemos que se muestren diferencias entre las distintas zonas en «Establecer color».

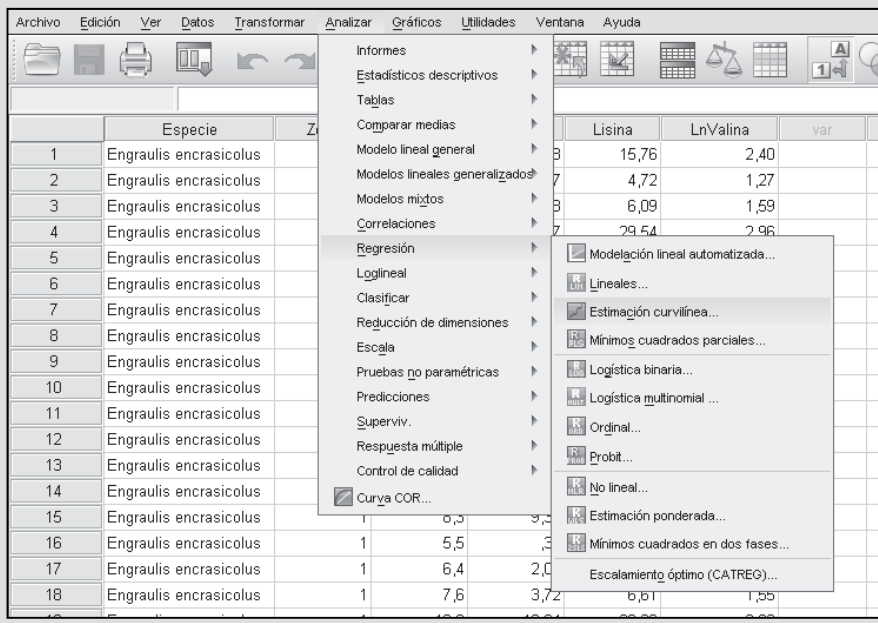


Paso 4. La gráfica que se obtiene parece mostrar una relación lineal, aunque es mejor realizar un análisis para comprobarlo.

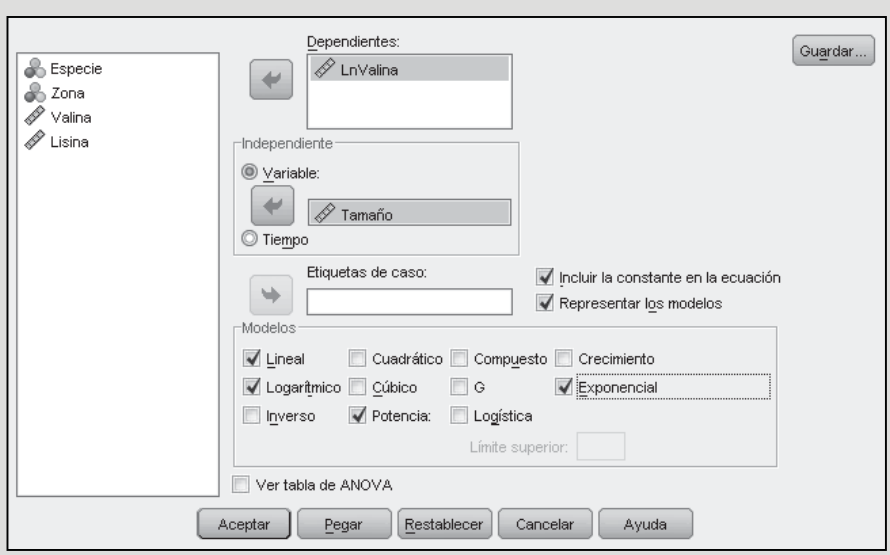


CUADRO IX.4. *(Continuación)*[SPSS]

Paso 5. Para ello vamos a «Analizar» ⇨ «Regresión» ⇨ «Estimación curvilínea…».



Paso 6. En «Dependientes» se selecciona «LnValina» y en «Variable» ponemos «Tamaño». En «Modelos» marcamos los modelos de regresión más frecuentes: «Lineal», «Logarítmico», «Potencia» y «Exponencial».



CUADRO IX.4. *(Continuación)*[SPSS]

Paso 7. Los resultados muestran que el r^2 es más alto en el caso de la regresión logarítmica (0,632), pero hay muy poca diferencia con la lineal (0,629). Por tanto, como la relación entre «LnValina» y «Tamaño» se puede considerar lineal no es necesario transformar. Si fuera necesario transformar, en la Tabla IX.1 (véase el paso 5 en la aplicación con STATISTICA) se muestran las transformaciones que hay que hacer para que la relación entre variables sea lineal.

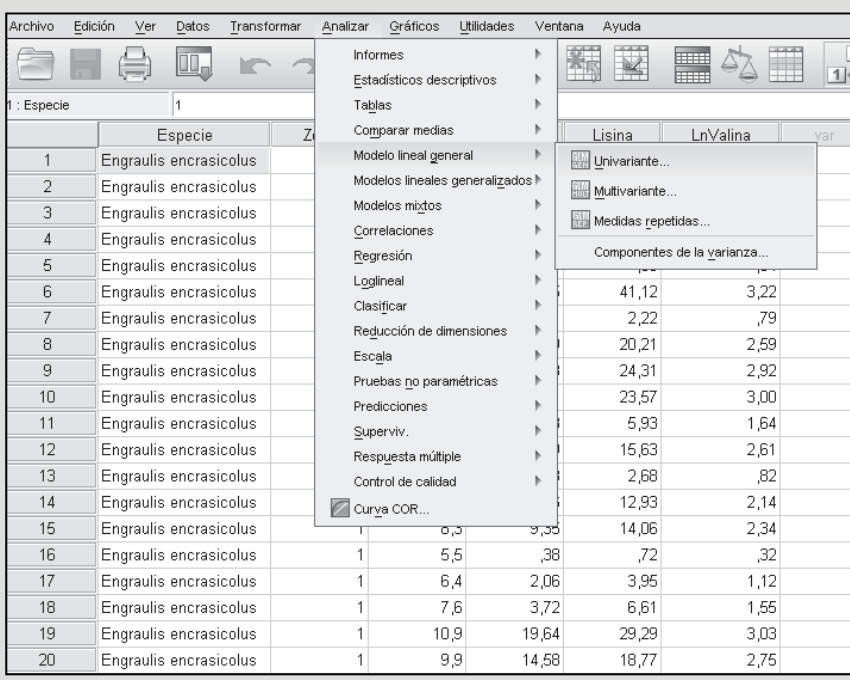
Resumen del modelo y estimaciones de los parámetros

Variable dependiente:LnValina

	Resumen del modelo					Estimaciones de los parámetros	
Ecuación	R cuadrado	F	gl1	gl2	Sig.	Constante	b1
Lineal	,629	399,899	1	236	,000	-,341	,237
Logarítmica	,632	404,916	1	236	,000	-2,460	1,974
Potencia	,597	349,558	1	236	,000	,054	1,562
Exponencial	,520	256,170	1	236	,000	,323	,175

La variable independiente esTamaño.

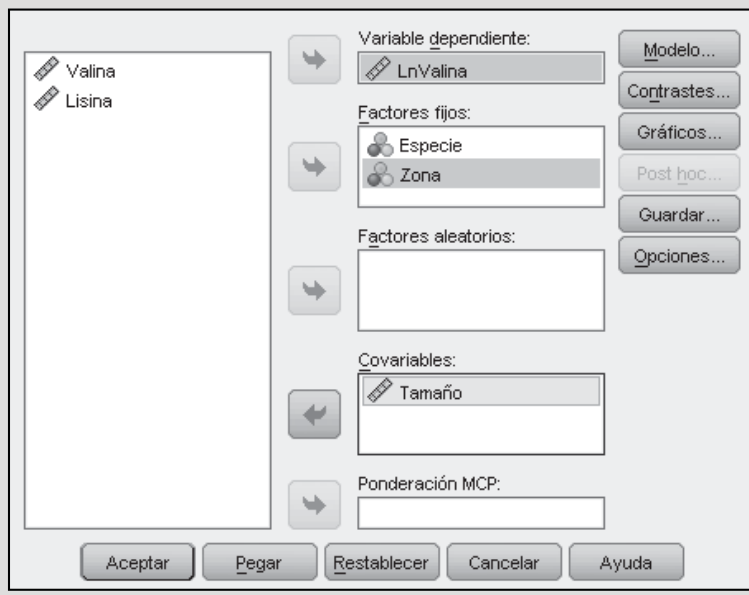
Paso 8. Para realizar el ANCOVA se selecciona «Analizar» ⇨ «Modelo lineal general» ⇨ «Univariante…». En el supuesto de querer analizar más de una variable dependiente, simplemente se seleccionaría «Multivariante…».



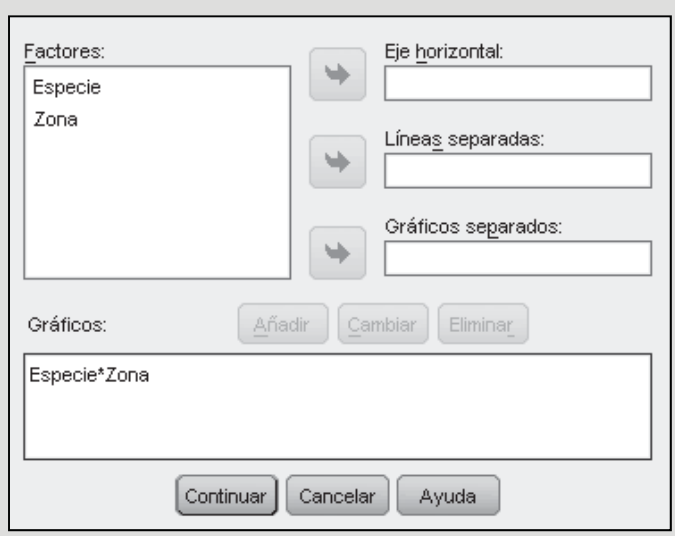
	Especie					
1	Engraulis encrasicolus					
2	Engraulis encrasicolus					
3	Engraulis encrasicolus					
4	Engraulis encrasicolus					
5	Engraulis encrasicolus					
6	Engraulis encrasicolus				41,12	3,22
7	Engraulis encrasicolus				2,22	,79
8	Engraulis encrasicolus				20,21	2,59
9	Engraulis encrasicolus				24,31	2,92
10	Engraulis encrasicolus				23,57	3,00
11	Engraulis encrasicolus				5,93	1,64
12	Engraulis encrasicolus				15,63	2,61
13	Engraulis encrasicolus				2,68	,82
14	Engraulis encrasicolus				12,93	2,14
15	Engraulis encrasicolus	1	8,3	9,35	14,06	2,34
16	Engraulis encrasicolus	1	5,5	,38	,72	,32
17	Engraulis encrasicolus	1	6,4	2,06	3,95	1,12
18	Engraulis encrasicolus	1	7,6	3,72	6,61	1,55
19	Engraulis encrasicolus	1	10,9	19,64	29,29	3,03
20	Engraulis encrasicolus	1	9,9	14,58	18,77	2,75

CUADRO IX.4. *(Continuación)*[SPSS]

Paso 9. En «Variable dependiente» se selecciona «LnValina», en «Factores fijos» se incluye «Especie» y «Zona», y en «Covariables» ponemos «Tamaño».

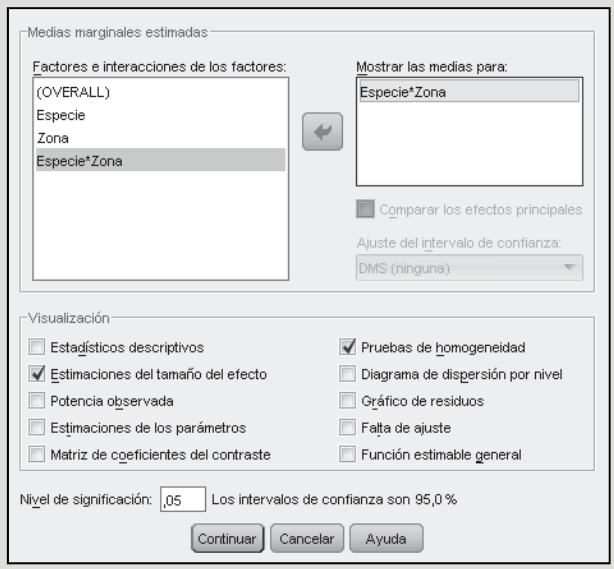


Paso 10. En un ANCOVA no se activa la opción de pruebas «Post hoc». En la opción «Gráficos…» aparecen los gráficos de perfil (gráficos de interacción), que sirven para comparar las medias marginales en el modelo (aumentan o disminuyen). Todos los factores fijos y aleatorios, si existen, están disponibles para los gráficos. Solo hay que seleccionar los factores que se quieren representar y pulsar «Añadir» para incluir los diferentes gráficos.

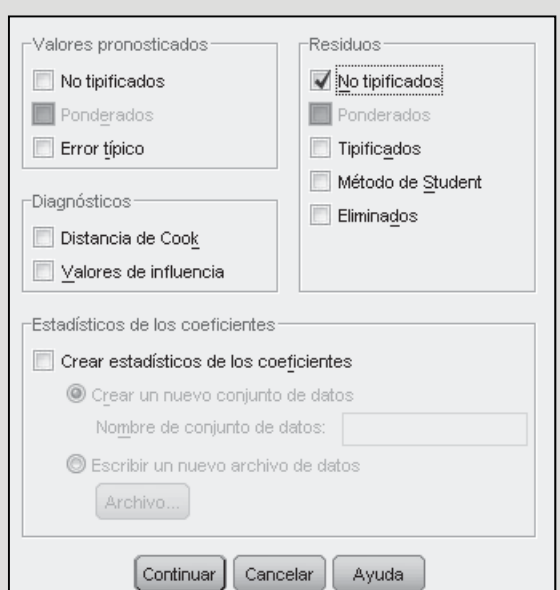


CUADRO IX.4. *(Continuación)*[SPSS]

Paso 11. En «Opciones» se marca «Estimaciones del tamaño del efecto», que permite cuantificar la importancia de cada factor sobre la variable dependiente por medio del estadístico Eta al cuadrado parcial, y «Pruebas de homogeneidad» para ver el requisito de homogeneidad de varianzas.



Paso 12. En la opción «Guardar» se marcan los residuos «No tipificados». Para despejar cualquier duda sobre el requisito de Normalidad, lo que se hace es generar los residuos y más tarde analizar si tienen una distribución Normal.



CUADRO IX.4. *(Continuación)*[SPSS]

Paso 13. Los resultados se muestran en la siguiente tabla. Existe una clara relación entre la concentración de Valina (en pg larva^{-1}) y el tamaño (en mm) ($F_{1,\ 231}$ = 380,7, p < 0,001), por lo tanto el efecto de la covariable es significativo. Se siguen sin observar diferencias significativas entre especies (ANOVA, $F_{1,\ 231}$ = 0,98, p = 0,323).

Pruebas de los efectos inter-sujetos

Variable dependiente:LnValina

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial
Modelo corregido	91,765[a]	6	15,294	72,462	,000	,653
Intersección	1,436	1	1,436	6,804	,010	,029
Tamaño	78,240	1	78,240	370,694	,000	,616
Especie	,207	1	,207	,982	,323	,004
Zona	1,339	2	,669	3,171	,044	,027
Especie * Zona	1,235	2	,618	2,926	,056	,025
Error	48,755	231	,211			
Total	898,032	238				
Total corregida	140,520	237				

a. R cuadrado = ,653 (R cuadrado corregida = ,644)

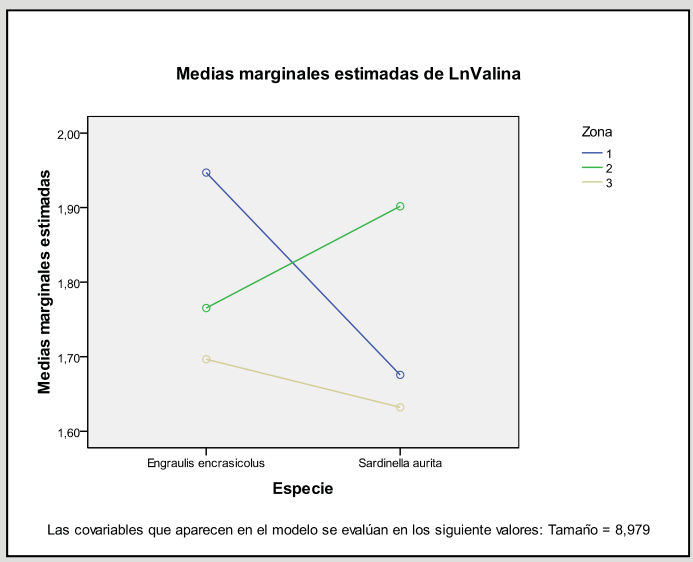
Existen diferencias significativas en la concentración del aminoácido entre zonas ($F_{2,\ 231}$ = 3,17, p < 0,044), pero es interesante mencionar que ahora, al considerar la covariable tamaño, estas diferencias no son tan importantes, ya que la probabilidad está muy cercana a 0,05. En el paso 33 del Cuadro IX.3 en la aplicación con SPSS las diferencias entre zonas eran más claras (ANOVA, $F_{2,\ 232}$ = 8,9, p < 0,001).

El efecto combinado de ambos factores («Especie» y «Zona») no lo podemos tener en cuenta, ya que el efecto individual del factor «Especie» sobre la variable dependiente «Valina» no es significativo, pero en cualquier caso no es significativo (p = 0,056).

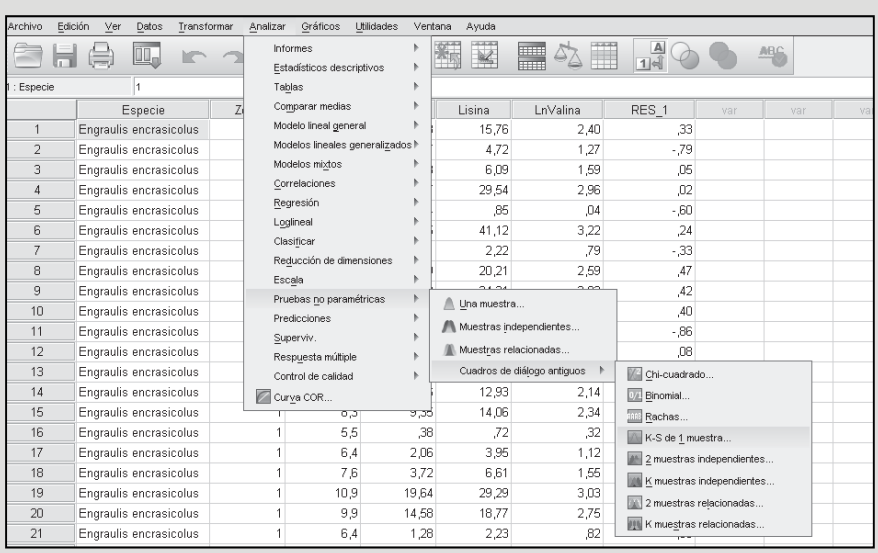
El estadístico Eta cuadrado parcial informa sobre el porcentaje de varianza residual de la variable dependiente que es explicada por cada factor, en un rango de 0 a 1, es decir la fracción de varianza no explicada ya por los restantes factores que puede ser explicada por ese factor. Se observa que la mayor influencia en la concentración de Valina la tiene el tamaño de la larva (0,616) y, aunque existen diferencias significativas en la concentración de Valina entre zonas, este factor no tiene una gran influencia en la concentración de Valina por larva (0,027).

CUADRO IX.4. *(Continuación)*[SPSS]

Paso 14. El gráfico de medias marginales permite visualizar la posible interacción entre los factores. El hecho de que las líneas se crucen parece indicar una interacción entre especie y zona, pero como el factor especie no es significativo, no procede considerar su interacción con otros factores; lo que observamos se debe simplemente al azar.



Paso 15. Para despejar cualquier duda con respecto al requisito de que los datos tengan una distribución Normal, analizamos la distribución de los residuos. Se selecciona «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «K-S de 1 muestra…».



CUADRO IX.4. *(Continuación)*[SPSS]

Paso 16. En «Lista Contrastar variables:» se introducen los residuos y en «Distribución de contraste» se marca «Normal».

Paso 17. El test de Kolmogorov-Smirnov muestra que los residuos tienen una distribución Normal (p = 0,502).

Prueba de Kolmogorov-Smirnov para una muestra

		Residuo para LnValina
N		238
Parámetros normales[a,,b]	Media	,0000
	Desviación típica	,45356
Diferencias más extremas	Absoluta	,054
	Positiva	,054
	Negativa	-,031
Z de Kolmogorov-Smirnov		,826
Sig. asintót. (bilateral)		,502

a. La distribución de contraste es la Normal.

Paso 18. Por último el test de Levene muestra que hay homogeneidad de varianzas (p = 0,315).

Contraste de Levene sobre la igualdad de las varianzas error[a]

Variable dependiente:LnValina

F	gl1	gl2	Sig.
1,189	5	232	,315

IX.1.6. ANÁLISIS DE VARIANZA Y COVARIANZA CON MEDIDAS REPETIDAS

En el apartado IX.1.3.2 se mostró que había una modalidad del *t*-test que se usaba en el caso de querer realizar un contraste de homogeneidad de datos relacionados o pareados. Sin embargo, se puede dar el caso de que el factor tenga más de dos niveles o que se tenga más de un factor. Por ejemplo, seleccionamos varias personas y les pedimos que den una puntuación de 0 a 10 en la calidad de cinco perfumes diferentes. En estos casos el *t*-test solo nos permite comparar dos perfumes entre sí; para comparar los cinco conjuntamente se utiliza el análisis de varianza para medidas repetidas.

El análisis con medidas repetidas está pensado para determinar con mayor precisión el error residual, que será menor si conseguimos controlar o eliminar la variabilidad entre sujetos y dejar únicamente la variabilidad intra sujetos. Cuanto más pequeño es el error residual, más precisión tendremos en la determinación de efectos significativos, ya que cada efecto se compara con la variabilidad residual para saber si es significativo.

Este tipo de análisis de varianza puede ser unifactorial o multifactorial, es decir, trabajar con uno o más factores con medidas repetidas en todos los factores o sólo en alguno de ellos. También permite incluir covariables, es decir, realizar un ANCOVA (un análisis de covarianza).

En el Cuadro IX.5 se muestran varios ejemplos de análisis de varianza con medidas repetidas

CUADRO IX.5. Análisis de varianza y covarianza con medidas repetidas

Aplicación con R.

EJEMPLO. En un experimento realizado con expertos catadores y personas que no tenían ninguna experiencia de cata, se les enseñó a identificar 15 tipos de vinos y 20 tipos de aceite de oliva de diferentes regiones. El objetivo era determinar si la capacidad de acertar la procedencia del vino y/o el aceite de oliva variaba en el tiempo (después de una hora, un día, una semana y un mes), entre los expertos y no expertos, y si la edad de las personas influía en la capacidad de acierto. Para ello, en cada tiempo cada persona evaluó un número grande de muestras y se registró el grado de acierto en una escala de 0 a 12.

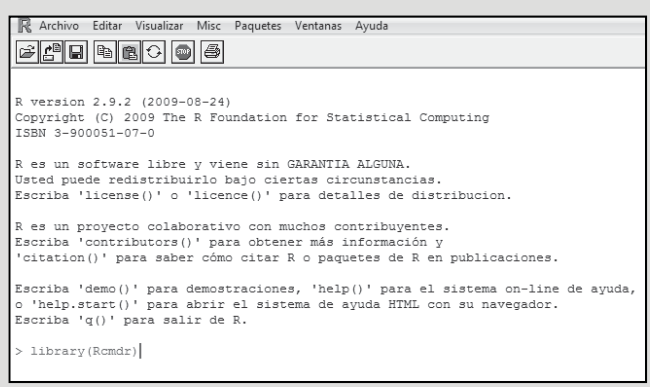
Realizaremos una aplicación de ANOVA unifactorial univariado con un factor con medidas repetidas, un ANOVA multifactorial en el cual hay factores con medidas repetidas (factores intra-sujetos) y factores sin medidas repetidas (factores inter-sujetos) y, por último, un ANCOVA con una variable cuantitativa independiente.

CUADRO IX.5. *(Continuación)*[R]

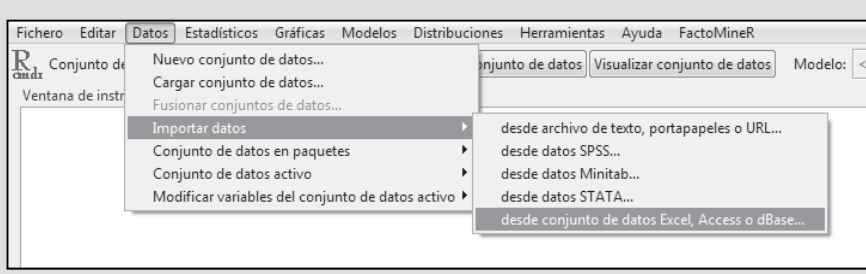
Los datos están en el archivo **Cuadro IX.5.xls.** Para la aplicación con R es conveniente que la variable dependiente «Y», en este caso el grado de acierto, esté en una única columna, con otras variables adicionales indicando cada factor: experiencia (SI/NO), edad, tiempo de la medida (H = Hora, D = Día, S = Semana, M = Mes) y –cuando existen repeticiones– el individuo. Otra variable indica si la muestra es de aceite de oliva (A) o de vino (V).

Individuo	Edad	Experiencia	Tiempo	Y	AV
1	46	SI	H	10,7	A
1	46	SI	D	7,1	A
1	46	SI	S	4,4	A
1	46	SI	M	4,1	A
2	41	SI	H	8,1	A
2	41	SI	D	7,4	A
2	41	SI	S	4,5	A
2	41	SI	M	3,2	A
3	47	SI	H	9,4	A
3	47	SI	D	6,5	A
3	47	SI	S	5,2	A
3	47	SI	M	4,6	A

Paso 1. Iniciamos R y cargamos el paquete «Rcmdr» introduciendo la orden «library(Rcmdr)». Si el paquete no está instalado, la instalación se realiza como se explica en el Apéndice II.

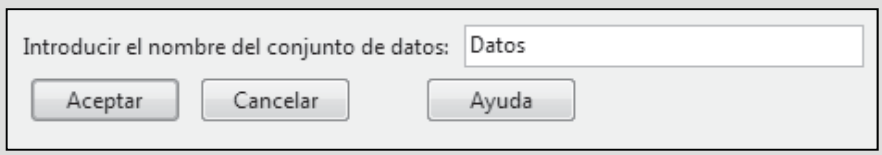

Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> library(Rcmdr)

Paso 2. Después de haber seleccionado el directorio de trabajo, como se explica en el Apéndice II, vamos a «Datos» ⇨ «Importar datos» ⇨ «desde conjunto de datos Excel, Access o dBase».


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda FactoMineR
Nuevo conjunto de datos...
Cargar conjunto de datos...
Fusionar conjuntos de datos...
Importar datos
Conjunto de datos en paquetes
Conjunto de datos activo
Modificar variables del conjunto de datos activo
desde archivo de texto, portapapeles o URL...
desde datos SPSS...
desde datos Minitab...
desde datos STATA...
desde conjunto de datos Excel, Access o dBase...
Visualizar conjunto de datos
Modelo:
Ventana de instr

CUADRO IX.5. *(Continuación)*[R]

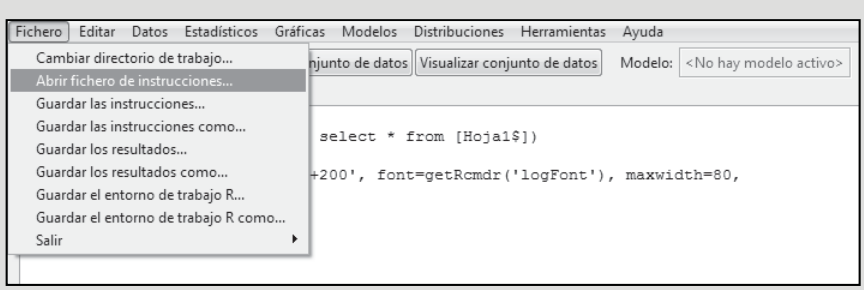
Paso 3. Aparece un pequeño cuadro de diálogo para introducir el nombre del conjunto de datos, en el que dejamos el nombre por defecto «Datos» y pulsamos «Aceptar». Posteriormente aparecen los archivos del directorio de trabajo, en el que elegimos **Cuadro IX.5.xls** y pulsamos «Abrir».



Paso 4. En R Commander pulsamos en «Visualizar conjunto de datos» lo que nos permite inspeccionar nuestros datos.

	Individuo	Edad	Experiencia	Tiempo	Y	AV
1	1	46	SI	H	10.7	A
2	1	46	SI	D	7.1	A
3	1	46	SI	S	4.4	A
4	1	46	SI	M	4.1	A
5	2	41	SI	H	8.1	A
6	2	41	SI	D	7.4	A
7	2	41	SI	S	4.5	A
8	2	41	SI	M	3.2	A
9	3	47	SI	H	9.4	A
10	3	47	SI	D	6.5	A
11	3	47	SI	S	5.2	A
12	3	47	SI	M	4.6	A
13	4	48	SI	H	10.5	A
14	4	48	SI	D	8.2	A
15	4	48	SI	S	5.2	A
16	4	48	SI	M	4.7	A
17	5	56	SI	H	7.2	A
18	5	56	SI	D	7.2	A
19	5	56	SI	S	5.1	A
20	5	56	SI	M	3.1	A
21	6	54	SI	H	10.2	A
22	6	54	SI	D	9.1	A
23	6	54	SI	S	8.4	A
24	6	54	SI	M	6.4	A
25	7	52	SI	H	10.3	A
26	7	52	SI	D	8.4	A
27	7	52	SI	S	8.5	A
28	7	52	SI	M	7.5	A
29	8	42	SI	H	9.3	A
30	8	42	SI	D	8.5	A

Paso 5. A continuación se carga el archivo de instrucciones **Cuadro IX.5.R** seleccionando «Fichero» ⇨ «Abrir fichero de instrucciones…» y no se guarda la salida.

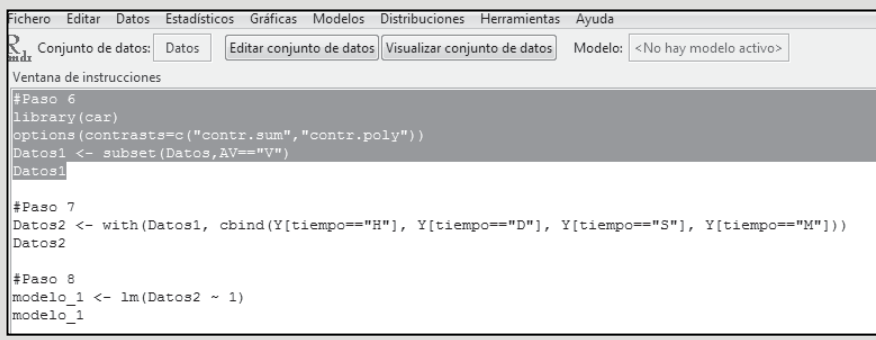


CUADRO IX.5. *(Continuación)*[R]

ANOVA unifactorial univariado con un factor con medidas repetidas

Para este primer ejemplo se va a estudiar si la capacidad de recordar la procedencia de un vino varía en el tiempo. Por lo tanto, tenemos un factor, que sería el tiempo, con cuatro niveles (una hora, un día, una semana y un mes) y la variable dependiente que es la capacidad de recordar la procedencia del vino.

Hay que ir seleccionando las diferentes instrucciones que se muestran en cada uno de los pasos y pulsar «Ejecutar». En la siguiente ventana, a modo de ejemplo se muestra el paso 6.



Paso 6. En primer lugar se carga el paquete "car" necesario para hacer los modelos lineales y luego se seleccionan los datos de vino.

library(car) # paquete de R con utilidades para modelos lineales
options(contrasts=c("contr.sum","contr.poly")) #suma cero en las ponderaciones de los efectos

Datos1 <- subset(Datos,AV=="V") # Selecciona los casos con AV = "V" (muestras de vino)
Datos1 # imprime en pantalla los datos seleccionados

```
> Datos1 <- subset(Datos,AV=="V")

> Datos1
    Individuo Edad Experiencia Tiempo    Y AV
77          1   46          SI      H 10.4  V
78          1   46          SI      D  9.2  V
79          1   46          SI      S  7.3  V
80          1   46          SI      M  6.2  V
81          2   41          SI      H 10.5  V
82          2   41          SI      D  9.2  V
83          2   41          SI      S  7.3  V
84          2   41          SI      M  5.2  V
85          3   47          SI      H 10.3  V
86          3   47          SI      D  8.1  V
```

CUADRO IX.5. *(Continuación)*[R]

Paso 7. Es necesario cambiar temporalmente el formato de los datos, creando un conjunto nuevo que llamaremos «Datos2», en el que solamente estarán los valores de la variable dependiente «Y» en una columna para cada período. Los datos están ordenados previamente por individuos, de modo que en el nuevo conjunto cada fila corresponde a un individuo o catador.

Datos2 <- with(Datos1, cbind(Y[Tiempo=="H"], Y[Tiempo=="D"], Y[Tiempo=="S"], Y[Tiempo=="M"])) # crea un conjunto de datos con las variables de medidas repetidas
Datos2 # imprime el nuevo conjunto de datos

```
> #Paso 7

> Datos2 <- with(Datos1, cbind(Y[Tiempo=="H"], Y[Tiempo=="D"], Y[Tiempo=="S"], Y[Tiempo=="M"]))

> Datos2
      [,1] [,2] [,3] [,4]
 [1,] 10.4  9.2  7.3  6.2
 [2,] 10.5  9.2  7.3  5.2
 [3,] 10.3  8.1  6.2  6.3
 [4,] 10.2  8.4  5.2  4.3
 [5,] 10.6  9.5  8.3  5.1
 [6,] 10.7  9.2  9.3  7.2
 [7,] 10.1  9.2  8.1  7.2
 [8,]  9.2  8.3  5.2  5.6
 [9,]  9.6  7.3  6.2  6.7
[10,] 10.7  9.2  9.6  9.1
[11,]  9.1  6.2  3.7  1.4
[12,]  9.4  4.3  3.1  1.5
[13,]  8.5  3.1  2.4  1.2
[14,]  8.2  3.1  2.5  1.2
[15,]  7.2  5.2  4.2  3.1
[16,]  8.3  5.5  6.2  4.4
[17,]  8.3  6.6  5.1  4.5
[18,]  8.1  5.9  6.4  4.3
[19,]  9.2  5.1  2.2  2.2
```

Paso 8. Construimos un modelo lineal sencillo, cuyos resultados muestran cómo la intensidad media, en una escala de 0 a 12, es menor cuanto más tiempo pasa, es decir, el grado de acierto a la hora de identificar el origen del vino es mayor cuando solo ha pasado una hora (9,4), que cuando ha pasado un mes (4,563).

modelo_1 <- lm(Datos2 ~ 1) # genera un modelo lineal simple
modelo_1 # imprime el modelo lineal; las estimaciones son las medias de cada variable

```
> #Paso 8

> modelo_1 <- lm(Datos2 ~ 1)

> modelo_1

Call:
lm(formula = Datos2 ~ 1)

Coefficients:
             [,1]   [,2]   [,3]   [,4]
(Intercept)  9.400  6.979  5.711  4.563
```

CUADRO IX.5. *(Continuación)*[R]

Paso 9. Se realiza el ANOVA para ver si, efectivamente, la diferencia observada en la capacidad de identificar el vino cambia significativamente en el tiempo.

```
f_1 <- factor(c("H","D","S","M")) #define factor: hora, dia, semana y mes
f_1 # imprime los niveles del factor
resultado_1 <- Anova(modelo_1, idata = data.frame(f_1), idesign = ~f_1, type="III") # anova con un factor intra sujetos (medidas repetidas)
summary(resultado_1, multivariate=FALSE) # imprime los resultados del anova
```

```
Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

               SS num Df Error SS den Df       F    Pr(>F)
(Intercept) 3374.2      1    238.5     18 254.710 4.547e-12 ***
f_1          245.2      3     56.9     54  77.633 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Mauchly Tests for Sphericity

    Test statistic p-value
f_1        0.54716 0.07335


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

     GG eps Pr(>F[GG])
f_1 0.70026  2.282e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    HF eps Pr(>F[HF])
f_1 0.7949  5.136e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los resultados indican que el tiempo es un factor significativo (p < 0,001). Las diferencias observadas en el grado de acierto a lo largo del tiempo no son debidas al azar, sino que la capacidad aprendida de identificar correctamente los vinos se pierde con el tiempo.

Un requisito del estadístico F en el ANOVA con medidas repetidas es que las varianzas de las diferencias entre cada dos niveles del factor intra-sujetos, el tiempo, deben ser iguales, y los niveles independientes entre sí. Este supuesto equivale a afirmar que la matriz de varianzas-covarianzas es proporcional a la matriz unidad (varianzas iguales y covarianzas nulas), o que la dispersión conjunta es esférica. La tabla muestra los resultados de la prueba de esfericidad de Mauchley. Como el valor p es mayor que 0,05 (p = 0,073) se acepta la hipótesis de esfericidad.

Si no se cumpliera el requisito de esfericidad (si el valor p fuese menor que 0,05), se debería utilizar alguna de las dos modificaciones que también se muestran: Greenhouse-Geisser, y Huynh-Feldt, para corregir los grados de libertad a la baja lo que permite ajustar (incrementar) el valor p del contraste ANOVA. En este caso ambos valores p corregidos ($2.282*10^{-14}$ y $5.136*10^{-16}$) son prácticamente cero, por lo que muestran también una relación significativa entre el tiempo y el acierto.

CUADRO IX.5. *(Continuación)*[R]

Paso 10. Una vez probada la relación significativa, debemos averiguar cuáles de los niveles son distintos (al menos dos lo son). Para ello podemos utilizar una prueba *t* con los dos niveles que queremos comparar. Por ejemplo, las instrucciones siguientes permiten conocer si existe diferencia entre los valores de «Y» al cabo de una hora y al cabo de un día.

```
contraste <- Datos2[,1]-Datos2[,2] # hora está en la columna 1 y día en la columna 2
t.test(contraste) # realiza la prueba t entre ambos niveles.
```

```
        One Sample t-test

data:  contraste
t = 7.2223, df = 18, p-value = 1.019e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.716780 3.125326
sample estimates:
mean of x
 2.421053
```

Como el valor $p < 0{,}01$, podemos afirmar que el grado de acierto es significativamente distinto entre hora y día. Si hacemos múltiples comparaciones como ésta, debemos recordar que es necesario utilizar algún tipo de protección, especialmente cuando se hacen muchas comparaciones, para evitar el aumento de la probabilidad de error de tipo I. Por ejemplo, podemos dividir el nivel de significación por el número de comparaciones (método Bonferroni). Por otra parte, si H y D son distintos, también lo son H y S o H y M, ya que las diferencias observadas en el valor medio (paso 8) son mayores, por lo que no es necesario hacer el contraste.

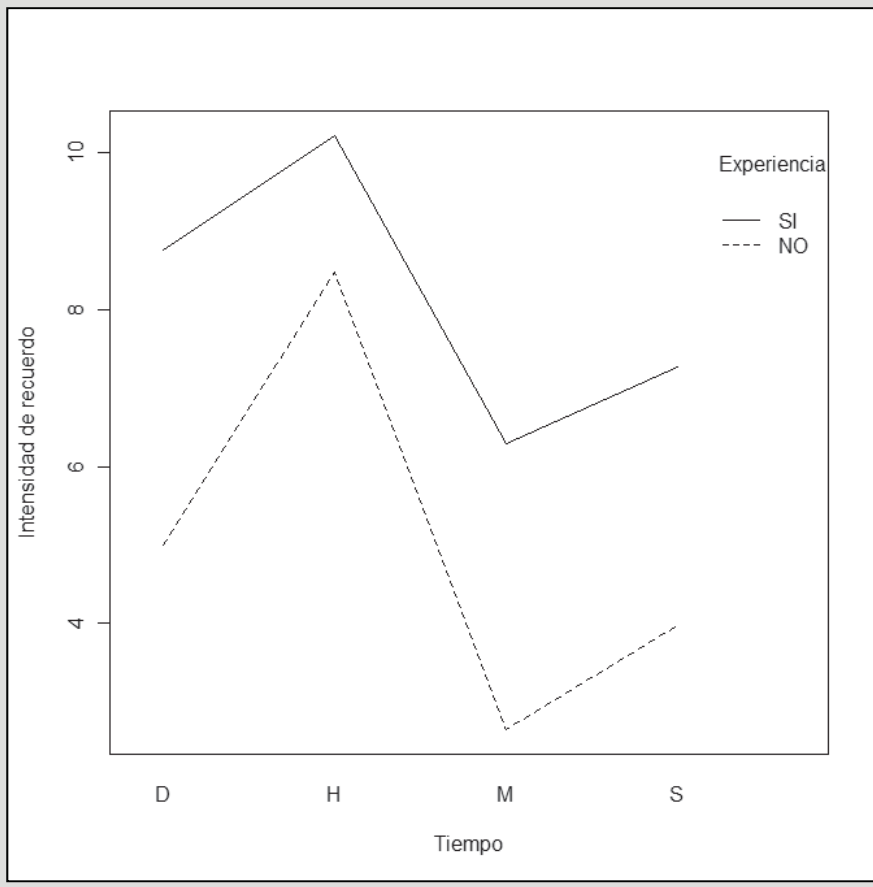
ANOVA multifactorial en el cual hay factores con medidas repetidas (factores intra-sujetos) y factores sin medidas repetidas (factores inter-sujetos)

En este caso analizaremos de nuevo si la capacidad de recordar la procedencia de un vino varía en el tiempo (factor intra-sujeto), pero analizando si existen diferencias entre expertos y no expertos catadores (factor inter-sujetos).

Paso 11. Comenzaremos construyendo un gráfico que ilustre la situación.

```
attach(Datos1) # utilizamos el conjunto Datos1, con la variable «Y» en una sola columna
interaction.plot(Tiempo, Experiencia, Y, trace.label = "Experiencia", xlab = "Tiempo", ylab = "Intensidad de recuerdo") # gráfico de interacción
```

CUADRO IX.5. *(Continuación)*[R]



Podemos observar como el grado de acierto, o intensidad del recuerdo, disminuye con el tiempo desde una hora hasta un mes, de forma similar en ambos grupos (ambas líneas son aproximadamente paralelas, especialmente entre un día y un mes). La intensidad es menor cuando el tiempo pasa, y los expertos mantienen siempre un grado de acierto dos o tres puntos más alto que los no expertos. Queremos ver si esta diferencia es estadísticamente significativa.

Paso 12. Realizamos el ANOVA con dos factores, además de la interacción de ambos.

```
res1 <- aov(Y ~ Experiencia*Tiempo + Error(factor(Individuo)))    # anova con dos factores
summary(res1) # imprime en pantalla los resultados
```

CUADRO IX.5. *(Continuación)*[R]

```
Error: factor(Individuo)
            Df  Sum Sq Mean Sq F value    Pr(>F)
Experiencia  1 183.556 183.556  56.843 8.105e-07 ***
Residuals   17  54.896   3.229
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
                  Df  Sum Sq Mean Sq F value    Pr(>F)
Tiempo             3 245.243  81.748 93.4530 < 2.2e-16 ***
Experiencia:Tiempo  3  12.250   4.083  4.6679  0.005876 **
Residuals          51  44.612   0.875
```

El factor «Experiencia» es significativo, con $p < 0{,}001$. El factor «Tiempo» también lo es ($p < 0{,}001$), como habíamos comprobado ya anteriormente, y también la interacción entre ambos, con $p = 0{,}005$, lo que significa que la evolución del recuerdo es distinta en ambos grupos, y las curvas no son realmente paralelas como nos había parecido antes. Por tanto, la capacidad de identificar el origen del vino disminuye con diferente intensidad a lo largo del tiempo entre los expertos y no expertos.

Análisis de covarianza (ANCOVA)

En el ejemplo anterior se observó que el grado de acierto al identificar la región de origen de un vino disminuía con el tiempo y que esta disminución era más acusada en las personas con menos experiencia en la cata de vinos. A continuación se quiere determinar si la «Edad» tiene alguna influencia en la capacidad para identificar vinos y su evolución en el tiempo.

Paso 13. La «Edad» es una variable cuantitativa, es decir una covariable, no un factor cualitativo, pero se puede incorporar al modelo de un modo sencillo.

```
res1 <- aov(Y ~ Experiencia*Tiempo + Edad + Error(factor(Individuo)))
# anova con un factor inter sujetos, otro intra sujetos (medidas repetidas) y una covariable
summary(res1)       # imprime los resultados en pantalla
```

```
Error: factor(Individuo)
            Df  Sum Sq Mean Sq F value    Pr(>F)
Experiencia  1 183.556 183.556 130.778  4.12e-09 ***
Edad         1  32.439  32.439  23.112 0.0001933 ***
Residuals   16  22.457   1.404
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
                  Df  Sum Sq Mean Sq F value    Pr(>F)
Tiempo             3 245.243  81.748 93.4530 < 2.2e-16 ***
Experiencia:Tiempo  3  12.250   4.083  4.6679  0.005876 **
Residuals          51  44.612   0.875
```

La edad también tiene un efecto significativo sobre el grado de acierto, con $p < 0{,}001$.

CUADRO IX.5. *(Continuación)*[STATISTICA]

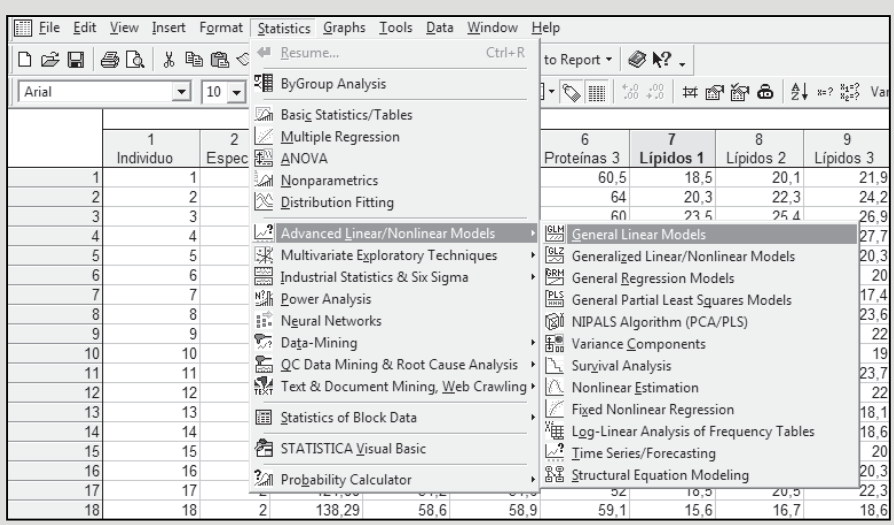
Aplicación con STATISTICA.

EJEMPLO. En el archivo **Cuadro IX.5.sta** hay datos del porcentaje de proteínas y lípidos medidos en dos especies de ganado vacuno, en tres semanas diferentes, así como el peso al nacer en kilogramos de los animales. El objetivo es determinar si existen diferencias en el tiempo en el porcentaje de proteínas y si estas posibles diferencias varían en función de la especie y del peso al nacer. Por lo tanto, la variable dependiente es el porcentaje de proteínas, el tiempo es un factor intra-sujetos (con medidas repetidas), la especie es un factor inter-sujetos (sin medidas repetidas) y el peso al nacer una covariable.

Se va a describir una aplicación de ANOVA unifactorial univariado con un factor con medidas repetidas, un ANOVA multifactorial en el cual hay factores con medidas repetidas (factores intra-sujetos) y factores sin medidas repetidas (factores inter-sujetos) y, por último, un ANCOVA.

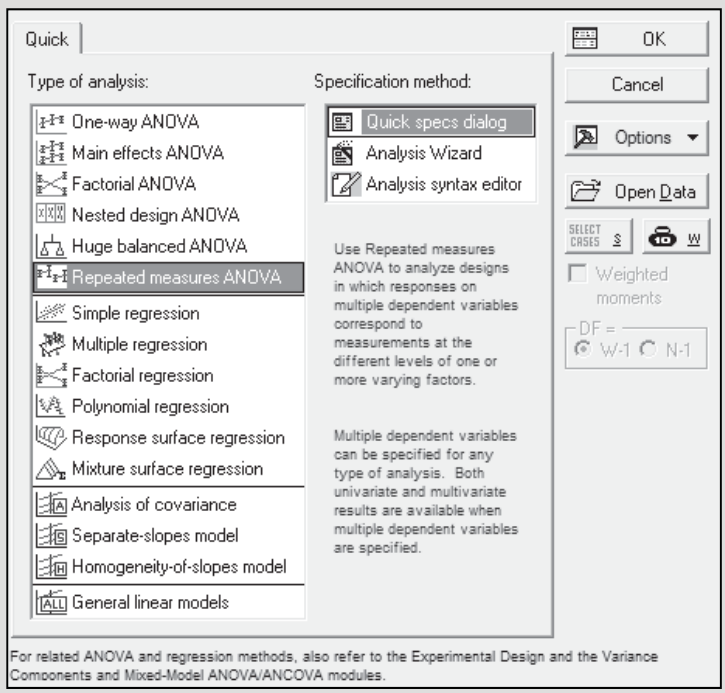
ANOVA unifactorial univariado con un factor con medidas repetidas

Paso 1. Para este primer ejemplo solo se analizará si el porcentaje de proteínas de los individuos varía en el tiempo. Por lo tanto, tenemos un factor intra-sujetos, que sería el tiempo, con tres niveles (cada una de las tres semanas analizadas) y la variable dependiente es el porcentaje de proteínas. Se selecciona «Statistics» ⇨ «Advanced Linear/Nonlinear Models» ⇨ «General Linear Models».

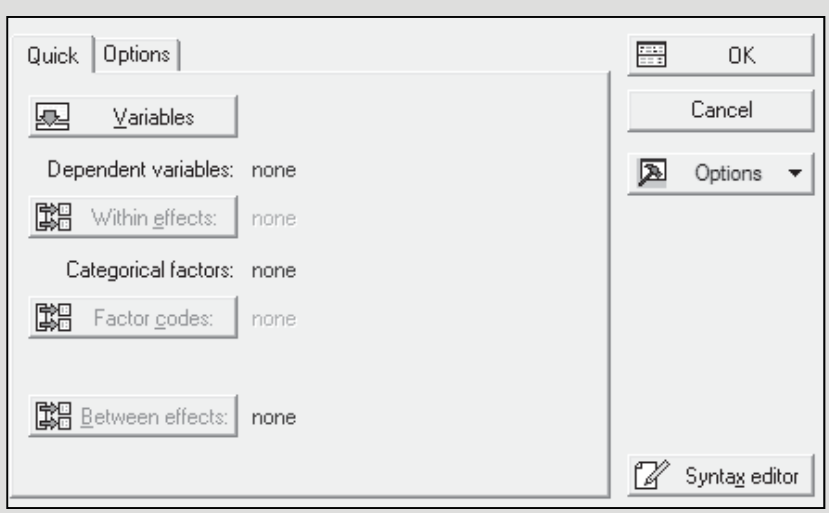


Paso 2. Aparece la siguiente ventana donde en «Type of analysis» se selecciona «Repeated measures ANOVA» y en «Specification method:» marcamos «Quick specs dialog».

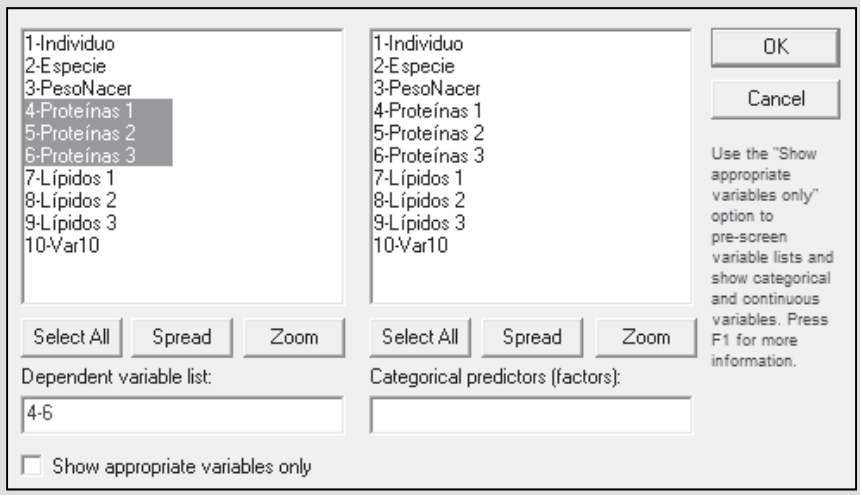
CUADRO IX.5. *(Continuación)* *STATISTICA*


Quick
Type of analysis:
One-way ANOVA
Main effects ANOVA
Factorial ANOVA
Nested design ANOVA
Huge balanced ANOVA
Repeated measures ANOVA
Simple regression
Multiple regression
Factorial regression
Polynomial regression
Response surface regression
Mixture surface regression
Analysis of covariance
Separate-slopes model
Homogeneity-of-slopes model
General linear models
Specification method:
Quick specs dialog
Analysis Wizard
Analysis syntax editor
Use Repeated measures ANOVA to analyze designs in which responses on multiple dependent variables correspond to measurements at the different levels of one or more varying factors.
Multiple dependent variables can be specified for any type of analysis. Both univariate and multivariate results are available when multiple dependent variables are specified.
OK
Cancel
Options
Open Data
Weighted moments
DF =
W-1
N-1
For related ANOVA and regression methods, also refer to the Experimental Design and the Variance Components and Mixed-Model ANOVA/ANCOVA modules.

Paso 3. En la siguiente ventana pulsamos en «Variables» y, en la ventana que aparece, en «Dependent variable list:» se incluyen los datos del porcentaje de proteínas de las tres semanas. Si además de incluir los datos de las proteínas también se incluyeran los de los lípidos, entonces estaríamos realizando un ANOVA unifactorial (un factor con medidas repetidas) multivariado (dos variables dependientes).

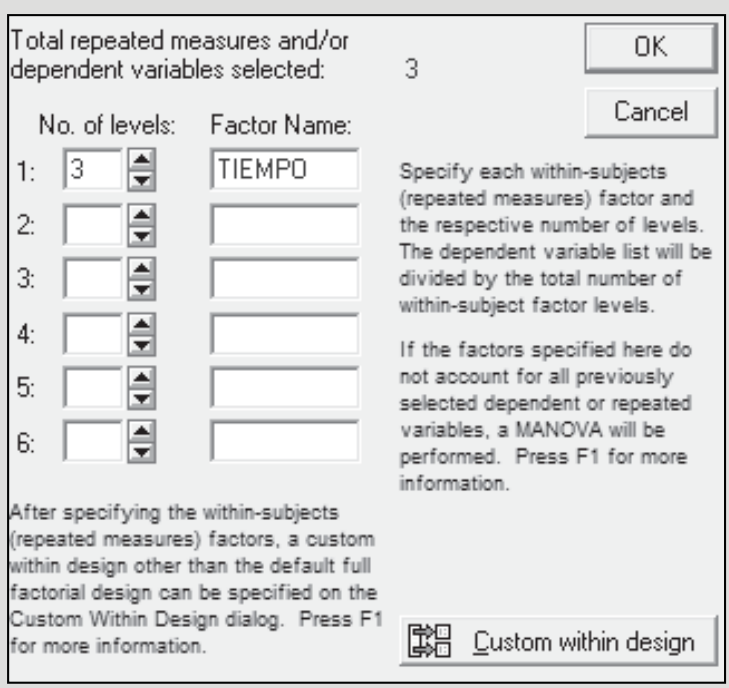


CUADRO IX.5. *(Continuación)* STATISTICA



Como en este primer caso no vamos a considerar el factor inter-sujetos, que sería el tipo de especie, no se incluye nada en «Categorical predictors (factors):» y pulsamos «OK».

Paso 4. En la ventana del paso 3 se pulsa en «Within effects:» y aparece un menú donde podemos incorporar todos los factores intra-sujetos. En nuestro análisis solo hay un factor intra-sujeto, el «Tiempo» que tiene 3 niveles, es decir, las tres semanas analizadas. Se pulsa «OK».

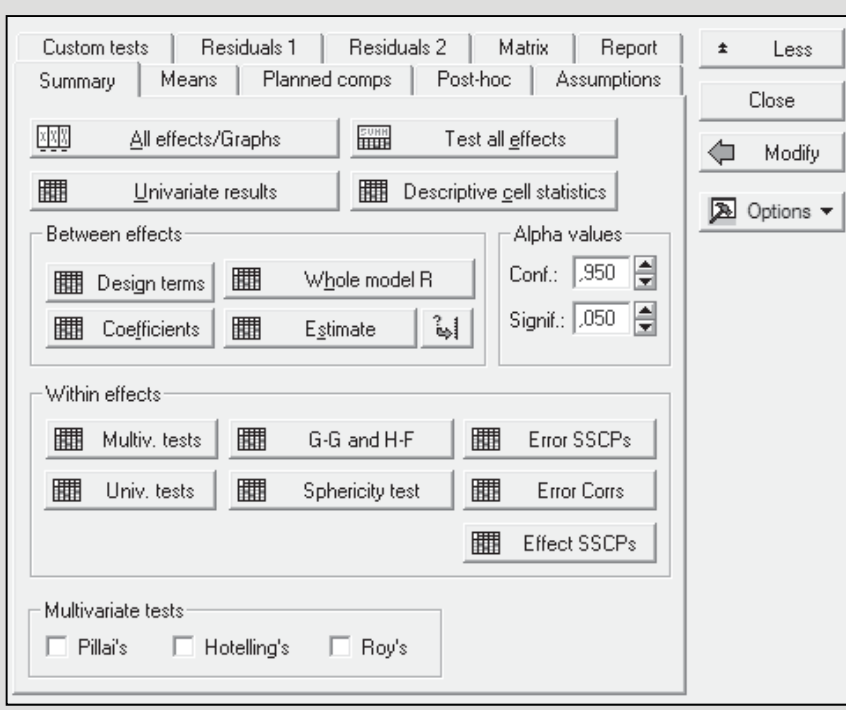

Total repeated measures and/or dependent variables selected: 3
OK
Cancel
No. of levels:
Factor Name:
1: 3 TIEMPO
2:
3:
4:
5:
6:
Specify each within-subjects (repeated measures) factor and the respective number of levels. The dependent variable list will be divided by the total number of within-subject factor levels.
If the factors specified here do not account for all previously selected dependent or repeated variables, a MANOVA will be performed. Press F1 for more information.
After specifying the within-subjects (repeated measures) factors, a custom within design other than the default full factorial design can be specified on the Custom Within Design dialog. Press F1 for more information.
Custom within design

CUADRO IX.5. *(Continuación)* *STATISTICA*

Si quisieramos introducir más factores con medidas repetidas, por ejemplo 3 factores, el primero con 2 niveles, el segundo también con dos niveles y el tercero con tres niveles, STATISTICA asigna los niveles a la variable dependiente comenzando por el último factor hasta el primero, como se muestra en la siguiente tabla. Por lo tanto, se deben colocar las columnas de la variable dependiente en la matriz de datos de la forma correcta para que la asignación entre los niveles de los factores intra-sujetos y la variable dependiente sea la adecuada.

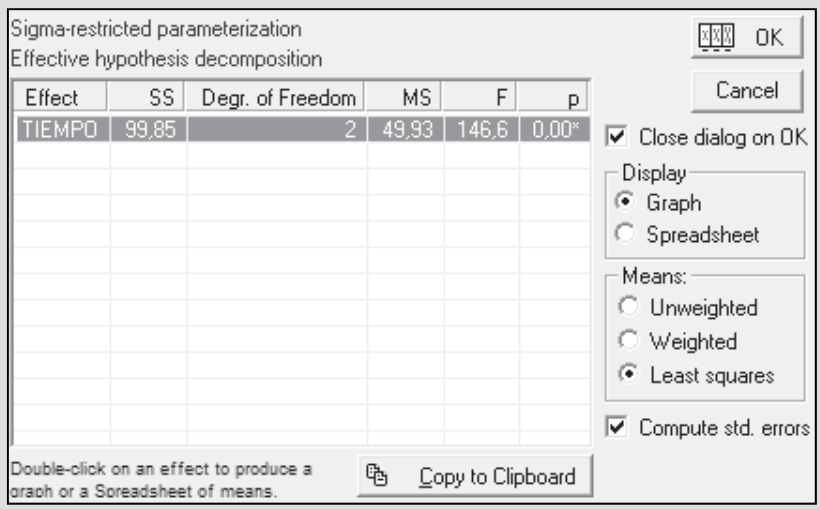
Factores con medidas repetidas	**Niveles de los factores con medidas repetidas**											
Primer factor	1	1	1	1	1	1	2	2	2	2	2	2
Segundo factor	1	1	1	2	2	2	1	1	1	2	2	2
Tercer factor	1	2	3	1	2	3	1	2	3	1	2	3
Nº de columna de la variable dependiente	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**

Paso 5. Pulsamos «OK» en la ventana del paso 3 y, en la ventana que aparece, pulsamos en «More results» para ampliar el número de diferentes opciones. Se selecciona la pestaña «Summary».



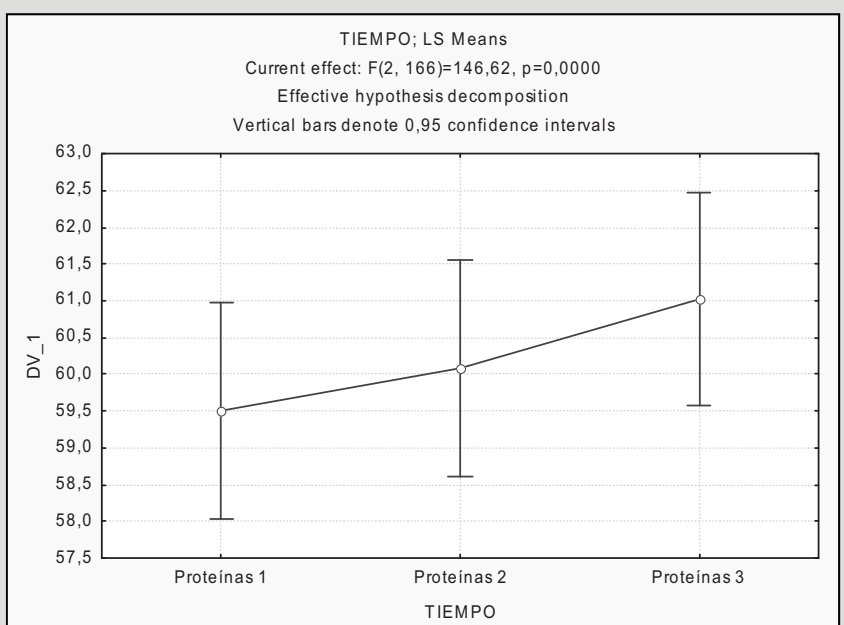
CUADRO IX.5. *(Continuación)*[STATISTICA]

Paso 6. Los resultados se pueden ver de dos formas diferentes. La primera es usando el estadístico *F* univariado. Para ello pulsamos en la ventana anterior en «All effects/Graphs». Se obtienen diferencias significativas en el porcentaje de proteínas en el tiempo ($F_{2,166}$ = 146,6, p < 0,001).



Effect	SS	Degr. of Freedom	MS	F	p
TIEMPO	99,85	2	49,93	146,6	0,00*

Paso 7. En la ventana anterior en «Display» se selecciona «Graph» y se pulsa «OK». El gráfico muestra como, en principio, el porcentaje de proteínas en los individuos parece aumentar en el tiempo.



CUADRO IX.5. *(Continuación)*[STATISTICA]

Paso 8. Un requisito del ANOVA con medidas repetidas es que las varianzas de las diferencias entre cada dos niveles del factor intra-sujetos, en este caso el tiempo, deben ser iguales, y los niveles independientes entre sí. Este supuesto equivale a afirmar que la matriz de varianzas-covarianzas es proporcional a la matriz unidad, o que la dispersión conjunta es esférica. En la ventana del paso 5 pulsamos en «Sphericity test». La siguiente tabla muestra los resultados de la prueba de esfericidad de Mauchley. Como la probabilidad es menor que 0,05 (p < 0,001) se rechaza en nuestro ejemplo la hipótesis de esfericidad.

	Mauchley Sphericity Test (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition			
Effect	W	Chi-Sqr.	df	p
TIEMPO	0,261017	110,1399	2	0,00

Paso 9. Cuando no se cumple el requisito de esfericidad (el valor p es menor que 0,05), es posible usar alguna modificación del test para corregir los grados de libertad a la baja lo que permite ajustar (incrementar) el valor p del contraste ANOVA. Para ello, en la ventana del paso 5 pulsamos en «G-G and H-F».

El estadístico épsilon (G-G épsilon, 0,575; H-F épsilon, 0,578) expresa el grado en que la matriz de varianzas covarianzas se aleja de la esfericidad: vale 1 en condiciones de esfericidad perfecta, y cuanto más pequeño mayor es la desviación de la esfericidad. Vemos que, después de las correcciones realizadas en los grados de libertad, en ambos casos p < 0,001 y por tanto se confirman las diferencias en el porcentaje de proteínas de los animales en el tiempo.

	Adjusted Univariate Tests for Repeated Measure: DV_1 (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition						
Effect	Degr. of Freedom	F	p	G-G Epsilon	G-G Adj. df1	G-G Adj. df2	G-G Adj. p
TIEMPO	2	146,6221	0,00	0,575049	1,150098	95,45810	0,00
Error	166						

	Adjusted Univariate Tests for Repeated Measure: DV_1 (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition						
Effect	Degr. of Freedom	F	p	H-F Epsilon	H-F Adj. df1	H-F Adj. df2	H-F Adj. p
TIEMPO	2	146,6221	0,00	0,577937	1,155874	95,93756	0,00
Error	166						

Una tercera corrección, que también podemos utilizar, se basa en el valor mínimo que puede tomar épsilon, que es igual al inverso del número de repeticiones menos uno (en este caso 1/(3-1) = 0,5). Con esta corrección, como se observa en la tabla siguiente (Lower Bound Adjusted P), se obtiene un valor p < 0,001 en el contraste, lo que corrobora las diferencias observadas en el porcentaje de proteínas a lo largo de las tres semanas.

CUADRO IX.5. *(Continuación)*[STATISTICA]

Effect	Adjusted Univariate Tests for Repeated Measure: DV_1 (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition						
	Degr. of Freedom	F	p	Lowr.Bnd Epsilon	Lowr.Bnd Adj. df1	Lowr.Bnd Adj. df2	Lowr.Bnd Adj. p
TIEMPO	2	146,6221	0,00	0,500000	1,000000	83,00000	0,00
Error	166						

Paso 10. Las correcciones mostradas en el punto anterior solo funcionan razonablemente bien cuando la desviación de la esfericidad es pequeña. Existe otro procedimiento alternativo a la corrección, que consiste en utilizar un contraste diferente que no requiera la hipótesis de esfericidad. Para ello, en la ventana del paso 5 pulsamos en «Multiv. tests» dentro del menú «Within tests». Se pueden usar diferentes estadísticos multivariados como Lambda de Wilks, Pillai, Hotelling y Roy los cuales deben ser seleccionados previamente en la parte inferior izquierda en la ventana del paso 5.

Effect	Multivariate tests for repeated measure: DV_1 (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition					
	Test	Value	F	Effect df	Error df	p
TIEMPO	Wilks	0,291002	99,89259	2	82	0,00

Estas pruebas multivariantes son menos potentes que el test F de medidas repetidas, pero deben ser utilizadas en lugar de las correcciones epsilon si la desviación de la esfericidad es importante, por ejemplo si epsilon es menor que 0,7. En esos casos, especialmente si el tamaño de muestra es grande (mayor que 10 veces el número de repeticiones), las pruebas multivariantes serán más potentes.

Paso 11. Para comparar dos a dos los niveles del factor tiempo, en la ventana del paso 5 se selecciona la pestaña «Post-hoc». Se puede usar cualquiera de las pruebas, pero una que se usa frecuentemente es «Bonferroni». Al pulsar el botón de esta prueba se obtienen los resultados que se muestran en la siguiente tabla. La comparación dos a dos de los niveles del factor muestra que también existen diferencias significativas ($p < 0,001$), en todos los casos, entre los diferentes tiempos, en el porcentaje de proteínas de los animales.

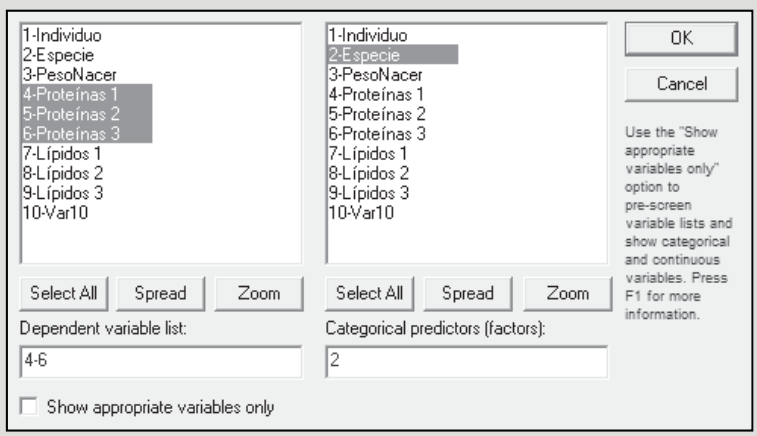
Cell No.	Bonferroni test; variable DV_1 (Cuadro IX.5.sta) Probabilities for Post Hoc Tests Error: Within MS = ,34050, df = 166,00			
	TIEMPO	{1} 59,507	{2} 60,073	{3} 61,032
1	Proteínas 1		0,000000	0,00
2	Proteínas 2	0,000000		0,00
3	Proteínas 3	0,000000	0,000000	

CUADRO IX.5. *(Continuación)*[STATISTICA]

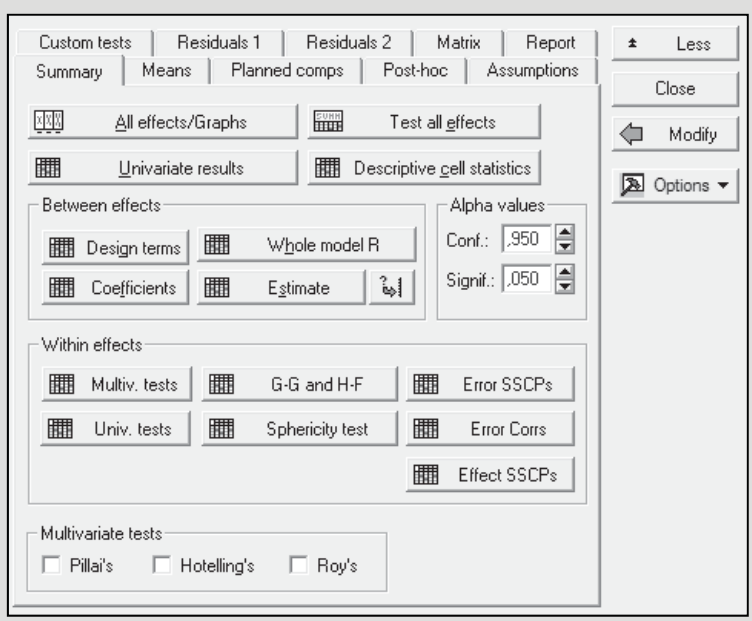
ANOVA multifactorial en el cual hay factores con medidas repetidas (factores intra-sujetos) y factores sin medidas repetidas (factores inter-sujetos)

En este caso analizaremos de nuevo si el porcentaje de proteínas cambia en el tiempo (factor intra-sujeto), pero considerando también la especie (factor inter-sujetos).

Paso 12. Se siguen los pasos 1 a 2. En «Dependent variable list:» se incluyen los datos del porcentaje de proteínas de las tres semanas y en «Categorical predictors (factors):» se selecciona el factor «Especie».

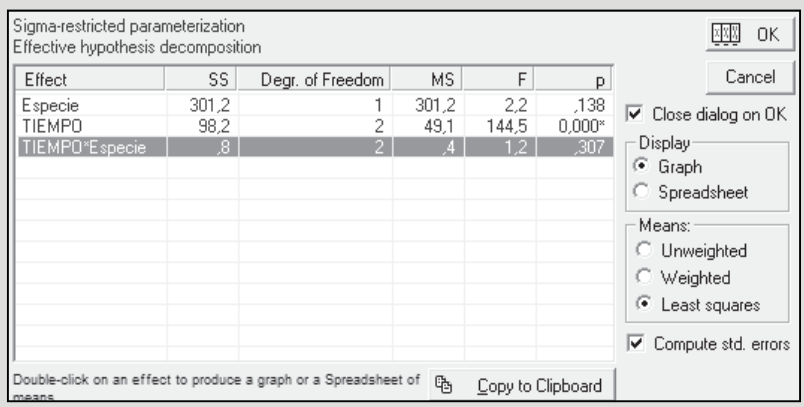


Paso 13. Se hace el mismo proceso que en el paso 4 para introducir el factor intra-sujetos «Tiempo». En la ventana que aparece, pulsamos en «More results» y se selecciona la pestaña «Summary».



CUADRO IX.5. *(Continuación)*[STATISTICA]

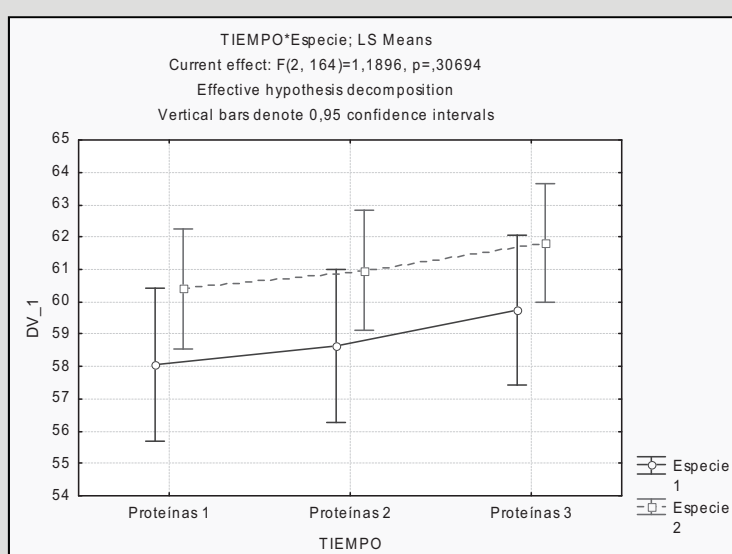
Paso 14. Para ver el estadístico *F* univariado pulsamos en la ventana anterior en «All effects/Graphs». No existen diferencias significativas en el porcentaje de proteínas entre especies ($F_{1,82}$ = 2,2, p = 0,138). Al añadir el factor «Especie», se siguen obteniendo diferencias significativas en el porcentaje de proteínas en el tiempo ($F_{2,164}$ = 144,5, p < 0,001). La interacción «Tiempo» y «Especie» no es significativa ($F_{2,164}$ = 1,2, p = 0,307), lo que indica que los cambios que se observan en el tiempo en el porcentaje de proteínas no son diferentes entre las dos especies. De todos modos si uno de los factores no es significativo (en este caso «Especie») no debe tenerse en cuenta la interacción aunque resulte significativa.


Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Especie	301,2	1	301,2	2,2	,138
TIEMPO	98,2	2	49,1	144,5	0,000*
TIEMPO*Especie	,8	2	,4	1,2	,307

OK
Cancel
Close dialog on OK
Display
Graph
Spreadsheet
Means:
Unweighted
Weighted
Least squares
Compute std. errors
Double-click on an effect to produce a graph or a Spreadsheet of means
Copy to Clipboard

Paso 15. En la ventana anterior en «Display» se selecciona «Graph» y se pulsa «OK». Hay que asegurarse de que esté seleccionada la fila «TIEMPO*Especie». Sale un menú donde indicamos «Tiempo» en el eje X y «Especie» como «Line pattern». El gráfico muestra que el porcentaje medio de proteínas sube de forma semejante en las dos especies, de ahí que la interacción no sea significativa.



CUADRO IX.5. *(Continuación)* STATISTICA

Paso 16. En la ventana del paso 13 pulsamos en «Sphericity test». Como la probabilidad es menor que 0,05 ($p < 0,001$) se rechaza la hipótesis de esfericidad.

	Mauchley Sphericity Test (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition			
Effect	W	Chi-Sqr.	df	p
TIEMPO	0,262767	108,2556	2	0,00

Paso 17. Como no se cumple el requisito de esfericidad, ya que la probabilidad es menor que 0,05, ya se mencionó que es posible usar cualquiera de los índices correctores épsilon. Para ello, en la ventana del paso 13 pulsamos en «G-G and H-F».

Se observa que, después de las correcciones realizadas por los épsilon, Greenhouse-Geisser (G-G épsilon, 0,576), Huynh-Feldt (H-F épsilon, 0,586) y Límite inferior (Lowr.Bd, 0,5), se obtiene una $p < 0,001$ para el tiempo y $p > 0,05$ para la interacción, en todos los casos. Por tanto, se confirman las diferencias en el porcentaje de proteínas de los animales en el tiempo, pero la interacción «Tiempo» y «Especie» no es significativa.

	Adjusted Univariate Tests for Repeated Measure: DV_1 (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition						
Effect	Degr. of Freedom	F	p	G-G Epsilon	G-G Adj. df1	G-G Adj. df2	G-G Adj. p
TIEMPO	2	144,4601	0,000000	0,575628	1,151256	94,40298	0,000000
TIEMPO*Especie	2	1,1896	0,306943	0,575628	1,151256	94,40298	0,285874
Error	164						

	Adjusted Univariate Tests for Repeated Measure: DV_1 (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition						
Effect	Degr. of Freedom	F	p	H-F Epsilon	H-F Adj. df1	H-F Adj. df2	H-F Adj. p
TIEMPO	2	144,4601	0,000000	0,585696	1,171391	96,05406	0,000000
TIEMPO*Especie	2	1,1896	0,306943	0,585696	1,171391	96,05406	0,286719
Error	164						

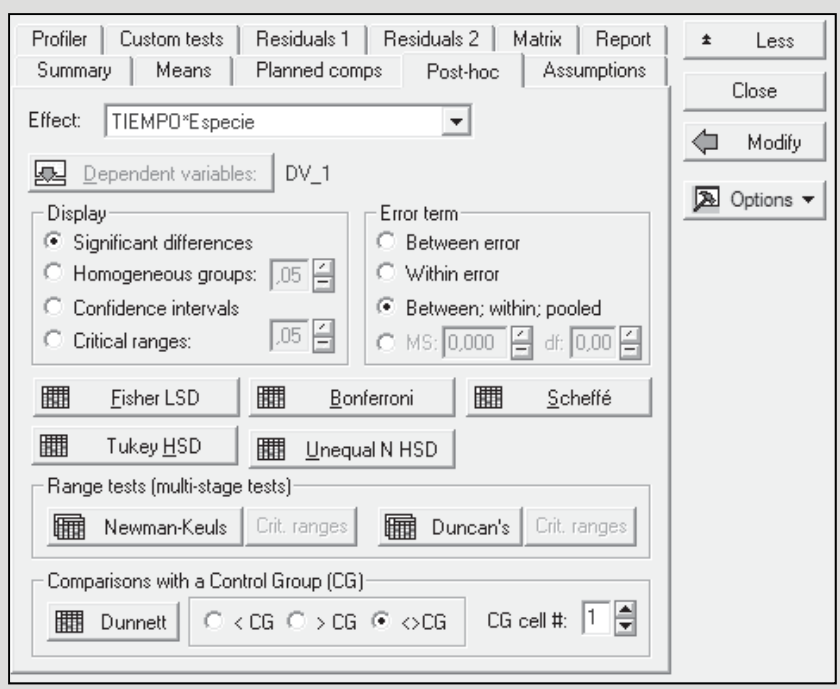
	Adjusted Univariate Tests for Repeated Measure: DV_1 (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition						
Effect	Degr. of Freedom	F	p	Lowr.Bnd Epsilon	Lowr.Bnd Adj. df1	Lowr.Bnd Adj. df2	Lowr.Bnd Adj. p
TIEMPO	2	144,4601	0,000000	0,500000	1,000000	82,00000	0,000000
TIEMPO*Especie	2	1,1896	0,306943	0,500000	1,000000	82,00000	0,278597
Error	164						

CUADRO IX.5. *(Continuación)*[STATISTICA]

Paso 18. Para ver el estadístico multivariado Lambda de Wilks, en la ventana del paso 13 pulsamos en «Multiv. tests» dentro del menú «Within tests». Los resultados confirman el cambio significativo en el tiempo de las proteínas ($p < 0{,}001$) y que estos cambios no son diferentes entre especies ($p = 0{,}408$).

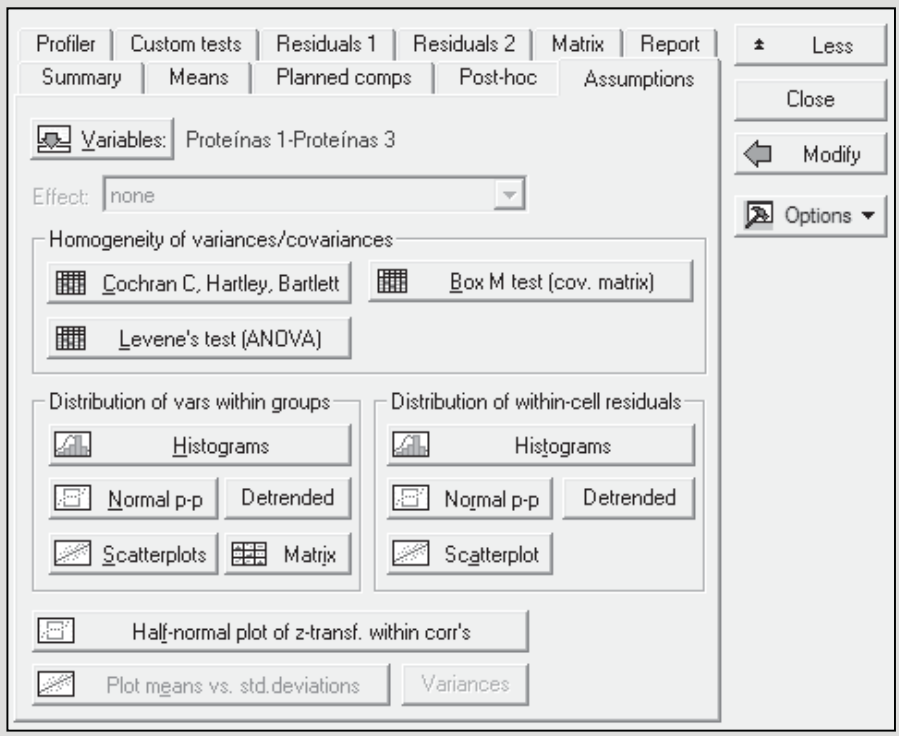
	Multivariate tests for repeated measure: DV_1 (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition					
Effect	Test	Value	F	Effect df	Error df	p
TIEMPO	Wilks	0,297805	95,49487	2	81	0,000000
TIEMPO*Especie	Wilks	0,978117	0,90610	2	81	0,408153

Paso 19. En el paso 11 ya se vio como poder comparar dos a dos los niveles del factor «Tiempo». Las pruebas *post-hoc* no tienen sentido con el factor «Especie» ya que hay solo dos niveles (dos especies). Aunque en la pestaña «Post-hoc», como vemos en la siguiente ventana, es posible hacer las pruebas *post-hoc* a la interacción de los factores, en este caso no tiene sentido ya que la interacción entre los factores no es significativa (véase paso 14).



Paso 20. Hay otro requisito que se debe cumplir para que los resultados tengan validez: la homogeneidad de varianzas entre los niveles del factor inter-sujeto. Para ello, seleccionamos la pestaña «Assumptions» y pulsamos «Box M test (cov. matrix)» y «Levene´s test (ANOVA)».

CUADRO IX.5. *(Continuación)*[STATISTICA]



El estadístico M de Box muestra que existe igualdad entre las matrices de varianzas-covarianzas para las dos especies (p = 0,333) y el test de Levene también muestra que existe homogeneidad de varianzas en el contenido de proteínas de cada semana entre las dos especies, ya que la probabilidad es siempre mayor que 0,05.

	Box M Test (Cuadro IX.5.sta) Effect: none (Computed for on all variables)			
	Box M	Chi-Sqr.	df	p
Box M	7,160917	6,853165	6	0,334640

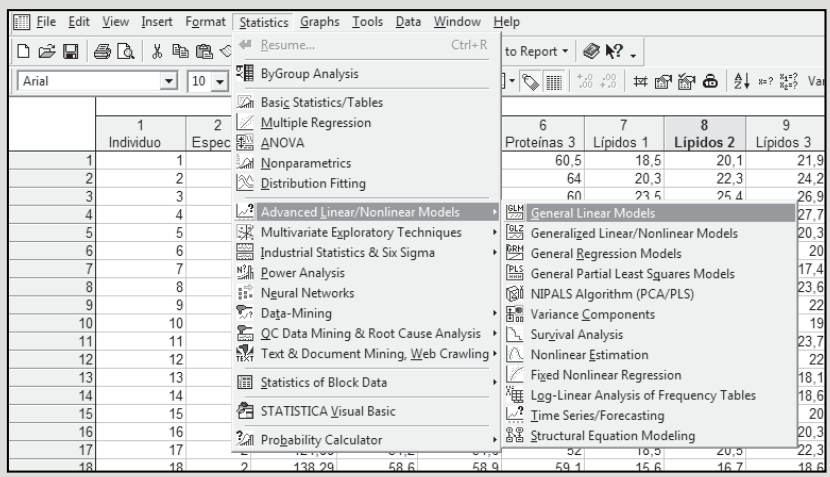
	Levene's Test for Homogeneity of Variances (Cuadro IX.5.sta) Effect: none Degrees of freedom for all F's: 1, 82			
	MS Effect	MS Error	F	p
Proteínas 1	22,99251	14,47051	1,588922	0,211055
Proteínas 2	26,86480	14,20543	1,891164	0,172816
Proteínas 3	28,36156	13,60945	2,083961	0,152663

CUADRO IX.5. *(Continuación)*[STATISTICA]

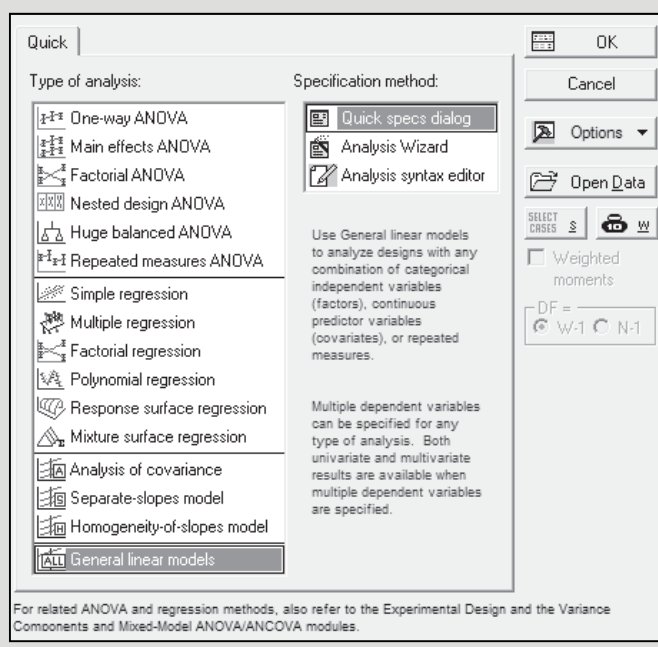
Análisis de covarianza (ANCOVA)

En último lugar, además de estudiar la variación del porcentaje de proteínas en el tiempo (factor intra-sujetos), en las dos especies (factor inter-sujetos), vamos analizar el posible efecto del peso al nacer (covariable).

Paso 21. Para realizar el ANCOVA se selecciona «Statistics» ⇨ «Advanced Linear/Nonlinear Models» ⇨ «General Linear Models».

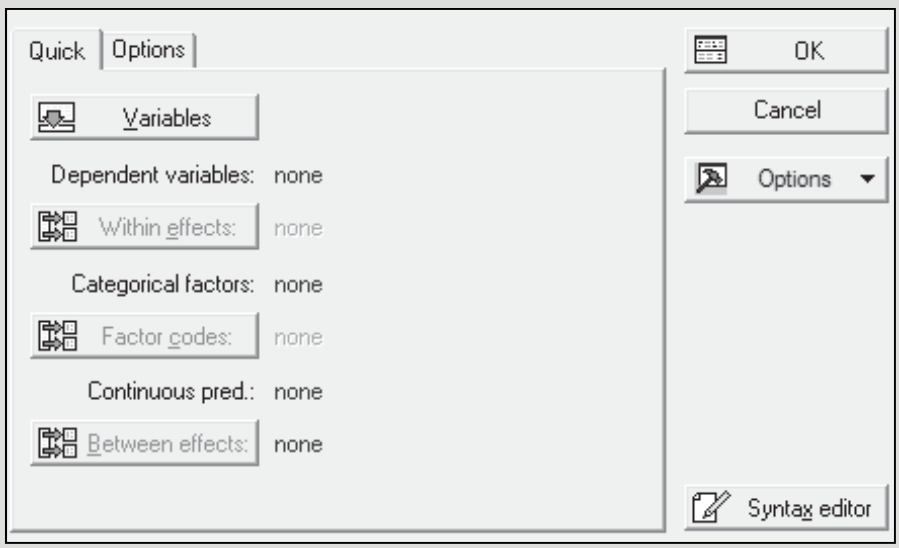


Paso 22. Posteriormente se selecciona «General linear models».

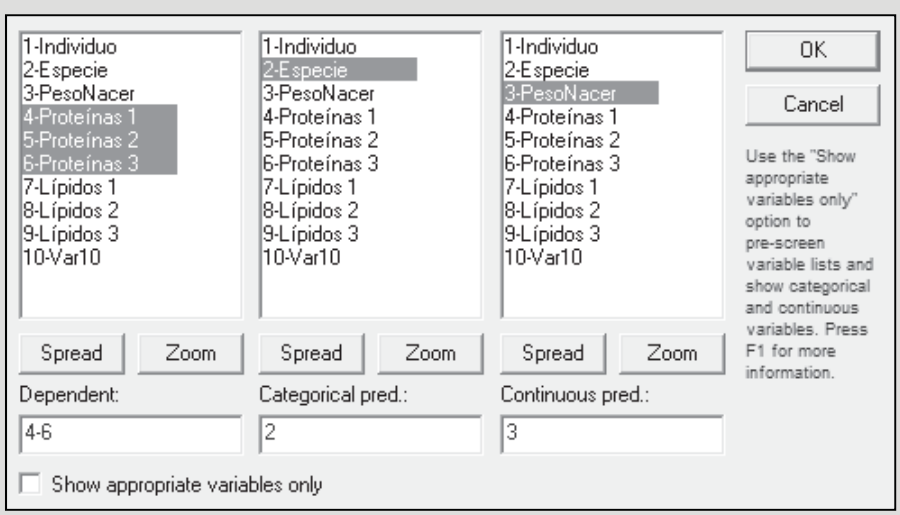

Quick
Type of analysis:
One-way ANOVA
Main effects ANOVA
Factorial ANOVA
Nested design ANOVA
Huge balanced ANOVA
Repeated measures ANOVA
Simple regression
Multiple regression
Factorial regression
Polynomial regression
Response surface regression
Mixture surface regression
Analysis of covariance
Separate-slopes model
Homogeneity-of-slopes model
General linear models
Specification method:
Quick specs dialog
Analysis Wizard
Analysis syntax editor
Use General linear models to analyze designs with any combination of categorical independent variables (factors), continuous predictor variables (covariates), or repeated measures.
Multiple dependent variables can be specified for any type of analysis. Both univariate and multivariate results are available when multiple dependent variables are specified.
OK
Cancel
Options
Open Data
Weighted moments
DF = W-1 N-1
For related ANOVA and regression methods, also refer to the Experimental Design and the Variance Components and Mixed-Model ANOVA/ANCOVA modules.

CUADRO IX.5. *(Continuación)*[STATISTICA]

Paso 23. En la ventana que aparece pulsamos en «Variables».



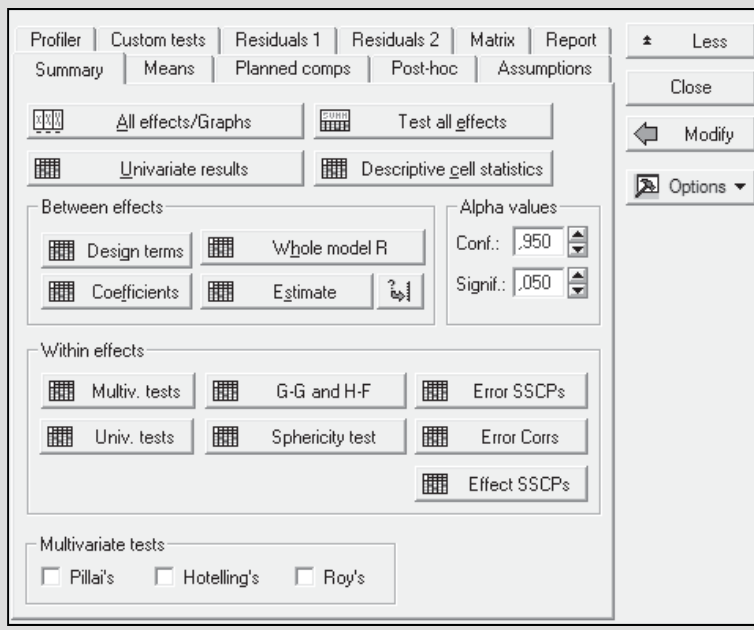
Paso 24. En «Dependent:» se incluyen los datos del porcentaje de proteínas de las tres semanas, en «Categorical pred.:» se selecciona el factor «Especie» y en «Continuous pred.:» se pone la covariable «PesoNacer».



Paso 25. Para introducir el factor intra-sujeto «Tiempo», se sigue lo descrito en el paso 4.

CUADRO IX.5. *(Continuación)*[STATISTICA]

Paso 26. En la ventana que aparece, pulsamos en «More results» para ampliar el número de diferentes opciones. Se selecciona la pestaña «Summary».



Paso 27. Para ver el estadístico *F* univariado pulsamos en la ventana anterior en «Test all effects». Se observan cambios importantes. Como la probabilidad es menor que 0,05, se rechaza la hipótesis de independencia y, por tanto, existe una fuerte relación entre el peso de los individuos al nacer y el porcentaje de proteínas ($F_{1,81}$ = 229,6, p < 0,001). El cambio más importante es que, al tener ahora en cuenta el peso al nacer, no existen diferencias significativas en el porcentaje de proteínas en el tiempo ($F_{2,162}$ = 0,05, p = 0,949). Por tanto, las diferencias en el porcentaje de proteínas en el tiempo observadas en los anteriores análisis, son en realidad debidas al diferente peso de los individuos al nacer.

	Repeated Measures Analysis of Variance (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition				
Effect	SS	Degr. of Freedom	MS	F	p
Intercept	638,065	1	638,065	18,0286	0,000058
PesoNacer	8125,111	1	8125,111	229,5762	0,000000
Especie	52,326	1	52,326	1,4785	0,227544
Error	2866,735	81	35,392		
TIEMPO	0,035	2	0,017	0,0514	0,949895
TIEMPO*PesoNacer	0,754	2	0,377	1,1111	0,331700
TIEMPO*Especie	0,965	2	0,482	1,4219	0,244260
Error	54,961	162	0,339		

CUADRO IX.5. *(Continuación)*[STATISTICA]

Al igual que ocurría en los otros análisis, no hay diferencias en el porcentaje de proteínas entre especies ($F_{1,81}$ = 1,47, p = 0,227). La interacción entre «Tiempo» y «Especie» sigue siendo no significativa ($F_{2,162}$ = 1,4, p = 0,244). Tampoco es significativa la interacción entre «Tiempo» y «PesoNacer» ($F_{2,162}$ = 1,1, p = 0,331), lo cual significa que la relación entre peso al nacer y porcentaje de proteínas no cambia en el tiempo. En todo caso no debe tenerse en cuenta una interacción cuando alguno de los factores involucrados en ella no es significativo.

Paso 28. En la ventana del paso 26 pulsamos en «Sphericity test». Como la probabilidad es menor que 0,05 (p < 0,001) se rechaza la hipótesis de esfericidad, que ya se mencionó es un requisito para poder usar el estadístico univariado.

	Mauchley Sphericity Test (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition			
Effect	W	Chi-Sqr.	df	p
TIEMPO	0,263541	106,6837	2	0,00

Paso 29. Las pruebas corregidas confirman lo mostrado anteriormente. Se observa que, después de las correcciones realizadas por los épsilon, Greenhouse-Geisser (G-G épsilon), Huynh-Feldt (H-F épsilon) y Límite inferior (Lowr.Bd), la p > 0,05 para el tiempo y p > 0,5 para las interacciones, en todos los casos.

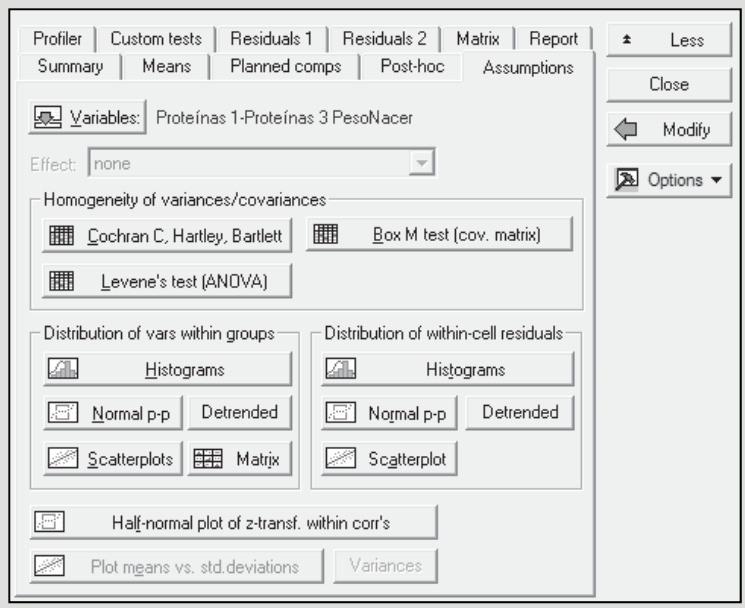
	Adjusted Univariate Tests for Repeated Measure: DV_1 (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition						
Effect	Degr. of Freedom	F	p	G-G Epsilon	G-G Adj. df1	G-G Adj. df2	G-G Adj. p
TIEMPO	2	0,051420	0,949895	0,575885	1,151769	93,29331	0,854127
TIEMPO*PesoNacer	2	1,111075	0,331700	0,575885	1,151769	93,29331	0,303783
TIEMPO*Especie	2	1,421857	0,244260	0,575885	1,151769	93,29331	0,240107
Error	162						

	Adjusted Univariate Tests for Repeated Measure: DV_1 (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition						
Effect	Degr. of Freedom	F	p	H-F Epsilon	H-F Adj. df1	H-F Adj. df2	H-F Adj. p
TIEMPO	2	0,051420	0,949895	0,593304	1,186609	96,11532	0,860690
TIEMPO*PesoNacer	2	1,111075	0,331700	0,593304	1,186609	96,11532	0,305558
TIEMPO*Especie	2	1,421857	0,244260	0,593304	1,186609	96,11532	0,240728
Error	162						

	Adjusted Univariate Tests for Repeated Measure: DV_1 (Cuadro IX.5.sta) Sigma-restricted parameterization Effective hypothesis decomposition						
Effect	Degr. of Freedom	F	p	Lowr.Bnd Epsilon	Lowr.Bnd Adj. df1	Lowr.Bnd Adj. df2	Lowr.Bnd Adj. p
TIEMPO	2	0,051420	0,949895	0,500000	1,000000	81,00000	0,821182
TIEMPO*PesoNacer	2	1,111075	0,331700	0,500000	1,000000	81,00000	0,294982
TIEMPO*Especie	2	1,421857	0,244260	0,500000	1,000000	81,00000	0,236580
Error	162						

CUADRO IX.5. *(Continuación)*[STATISTICA]

Paso 30. Como se mencionó anteriormente debe existir homogeneidad de varianzas entre los niveles del factor inter-sujeto y entre los niveles de la covariable. Para ello, seleccionamos la pestaña «Assumptions» y pulsamos «Box M test (cov. matrix)» y «Levene´s test (ANOVA)».



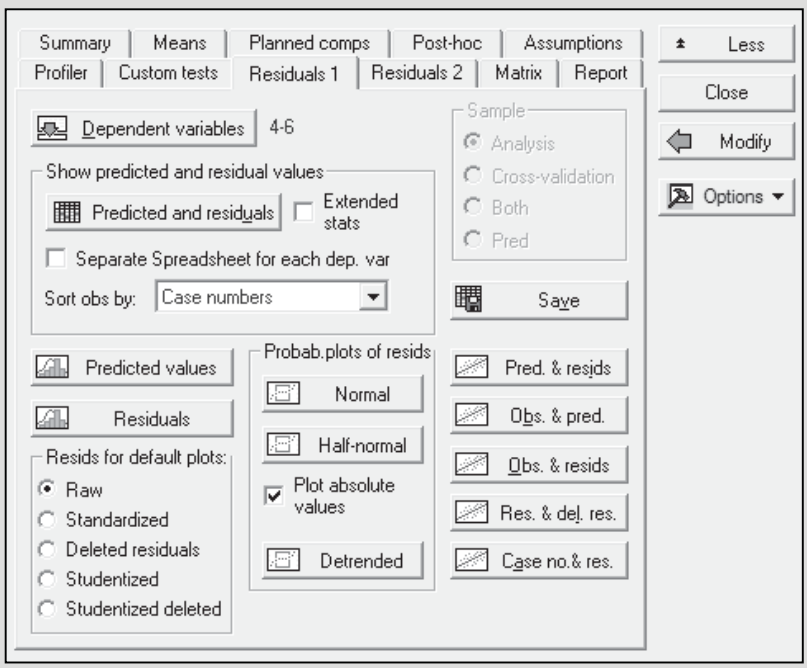
El estadístico M de Box muestra que no existe igualdad entre las matrices de varianzas-covarianzas (p = 0,045) y el test de Levene muestra que existe homogeneidad de varianzas en el contenido de proteínas de cada semana entre las dos especies y en función del peso al nacer, ya que la probabilidad es siempre mayor que 0,05 en todos los casos. Como ambas pruebas miden lo mismo, y el valor p de la M de Box es próximo a 0,05, la ligera desviación de la homogeneidad no debe afectar a la validez de los resultados.

	Box M Test (Cuadro IX.5.sta) Effect: none (Computed for all vars and covars)			
	Box M	Chi-Sqr.	df	p
Box M	19,74226	18,61968	10	0,045368

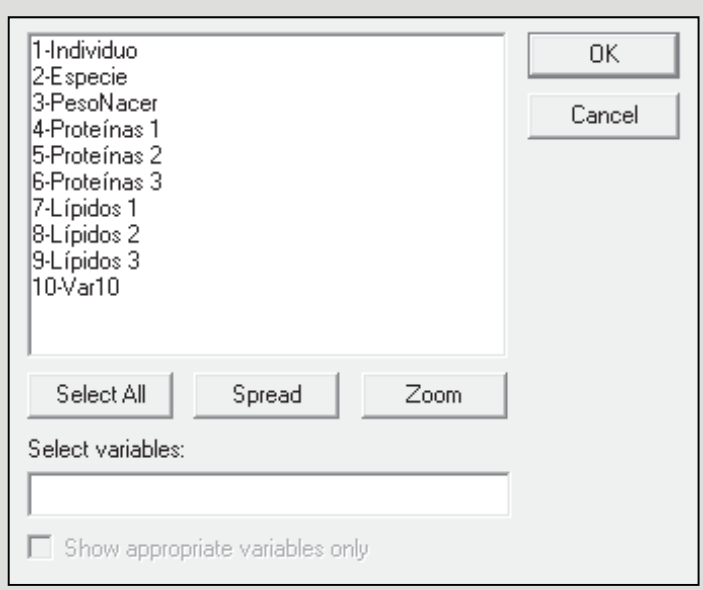
	Levene's Test for Homogeneity of Variances (Cuadro IX.5.sta) Effect: none Degrees of freedom for all F's: 1, 82			
	MS Effect	MS Error	F	p
Proteínas 1	22,99251	14,47051	1,588922	0,211055
Proteínas 2	26,86480	14,20543	1,891164	0,172816
Proteínas 3	28,36156	13,60945	2,083961	0,152663
PesoNacer	21,79717	29,68881	0,734188	0,394026

CUADRO IX.5. *(Continuación)* *STATISTICA*

Paso 31. Para solventar cualquier duda con respecto al requisito de distribución Normal, es conveniente ver si los residuos tienen distribución Normal. En la pestaña «Residuals 1» pulsamos en «Save» para guardar los residuos.



Paso 32. En la ventana que aparece pulsamos «OK» sin seleccionar ninguna variable.

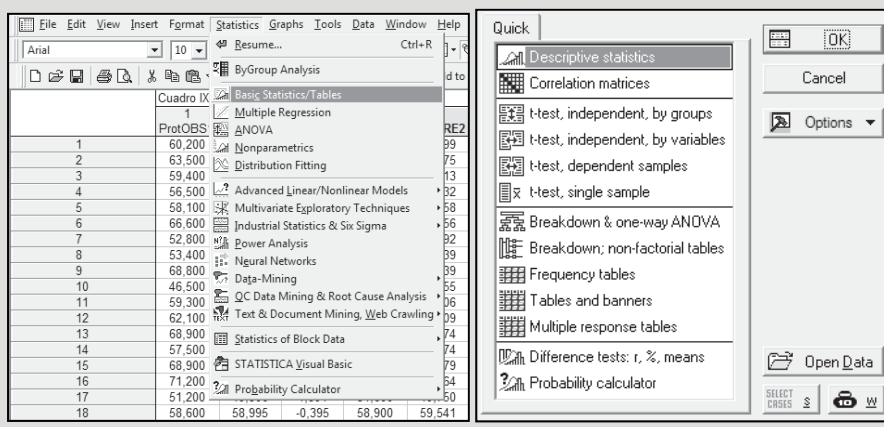


CUADRO IX.5. *(Continuación)*[STATISTICA]

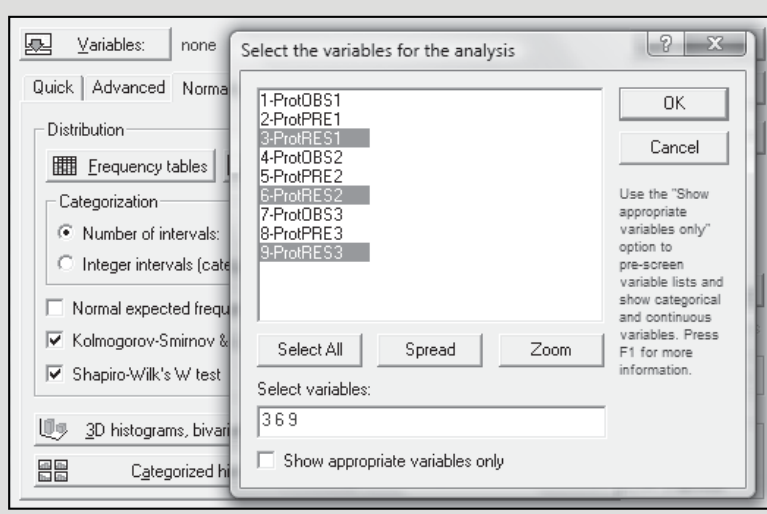
Paso 33. Se obtiene una tabla, en la cual los residuos de «Proteínas 1» son «ProtRES1», de «Proteínas 2» son «ProtRES2» y de «Proteínas 3» son «ProtRES3»

	Cuadro IX.5.sta								
	1 ProtOBS1	2 ProtPRE1	3 ProtRES1	4 ProtOBS2	5 ProtPRE2	6 ProtRES2	7 ProtOBS3	8 ProtPRE3	9 ProtRES3
1	60,200	57,830	2,370	60,300	58,399	1,901	60,500	59,506	0,994
2	63,500	61,071	2,429	63,700	61,675	2,025	64,000	62,824	1,176
3	59,400	57,151	2,249	59,500	57,713	1,787	60,000	58,811	1,189
4	56,500	53,113	3,387	56,900	53,632	3,268	57,300	54,676	2,624
5	58,100	57,393	0,707	58,400	57,958	0,442	58,500	59,059	-0,559
6	66,600	62,636	3,964	67,100	63,256	3,844	67,300	64,426	2,874
7	52,800	49,907	2,893	53,400	50,392	3,008	53,500	51,394	2,106
8	53,400	51,932	1,468	54,200	52,439	1,761	54,600	53,467	1,133

Paso 34. Se selecciona «Statistics» ⇨ «Basic Statistics/Tables» ⇨ «Descriptive statistics».

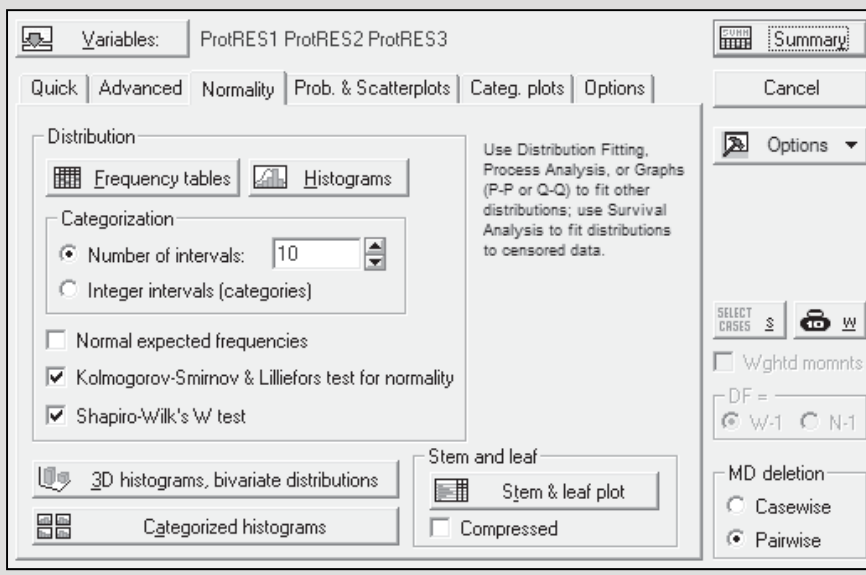


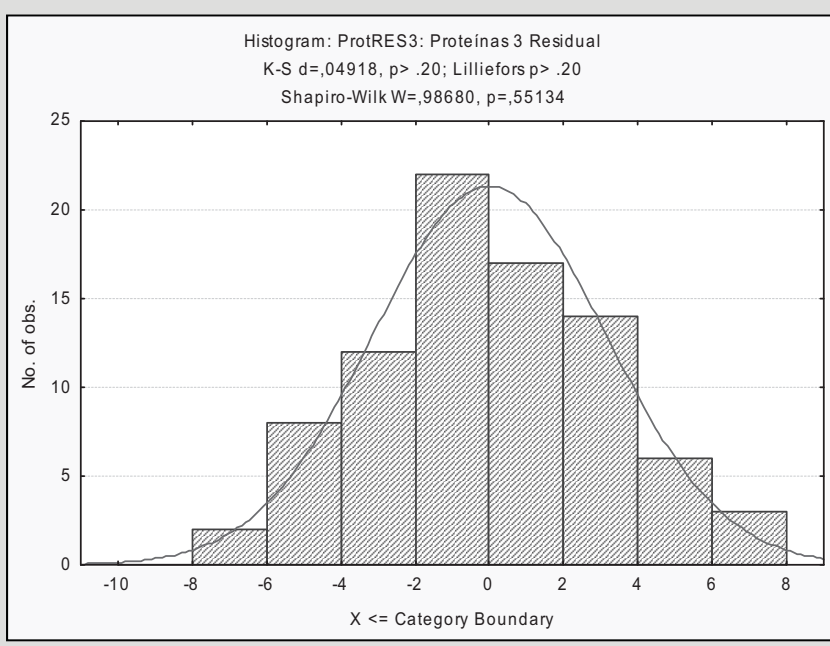
Paso 35. En la pestaña «Normality» marcamos los test «Kolmogorov-Smirnov & Lilliefors test for normality» y «Shapiro-Wilk´s test». En «Variables» se selecciona «ProtRES1», «ProtRES2» y «ProtRES3».



CUADRO IX.5. *(Continuación)*[STATISTICA]

Paso 36. En la siguiente ventana al pulsar «Histograms» aparecen tres gráficos (solo se muestra para el caso de «Proteínas 3»), donde se observa que tanto el test Kolmogorov–Smirnov con la corrección de Lilliefors, como el de Shapiro-Wilk muestran un $p > 0{,}05$ y, por tanto, se acepta la hipótesis nula de que los residuos se ajustan a una distribución Normal para «Proteínas 1», «Proteínas 2» y «Proteínas 3».





CUADRO IX.5. *(Continuación)*[SPSS]

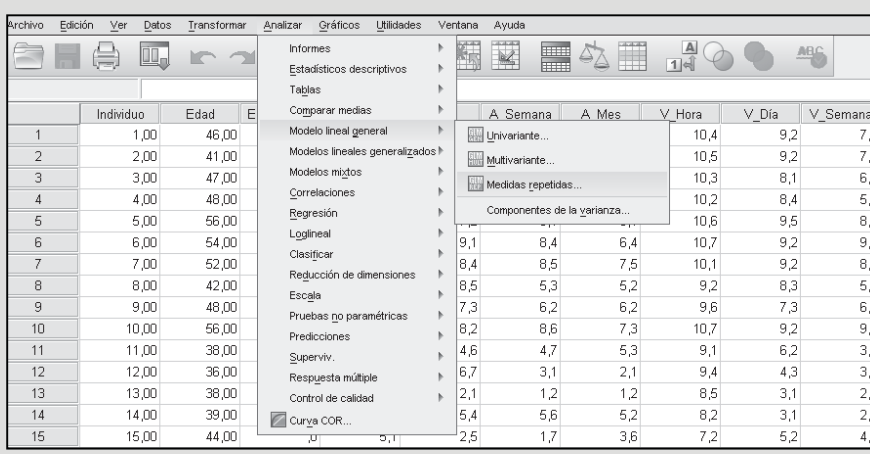
Aplicación con SPSS.

EJEMPLO. En un experimento realizado con expertos catadores y personas que no tenían ninguna experiencia de cata, se les enseñó a identificar 15 tipos de vino y 20 tipos de aceite de oliva de diferentes regiones. El objetivo era determinar si la capacidad de acertar la procedencia del vino y/o el aceite de oliva variaba en el tiempo (después de una hora, un día, una semana y un mes), entre los expertos y no expertos, y si la edad de las personas influía en la capacidad de acierto. Para ello, en cada tiempo cada persona evaluó un número grande de muestras y se registró el grado de acierto en una escala de 0 a 12. Los datos están en el archivo **Cuadro IX.5.V.sav.**

Se va a describir una aplicación de ANOVA unifactorial univariado con un factor con medidas repetidas, un ANOVA multifactorial en el cual hay factores con medidas repetidas (factores intra-sujetos) y factores sin medidas repetidas (factores inter-sujetos) y, por último, un ANCOVA.

ANOVA unifactorial univariado con un factor con medidas repetidas

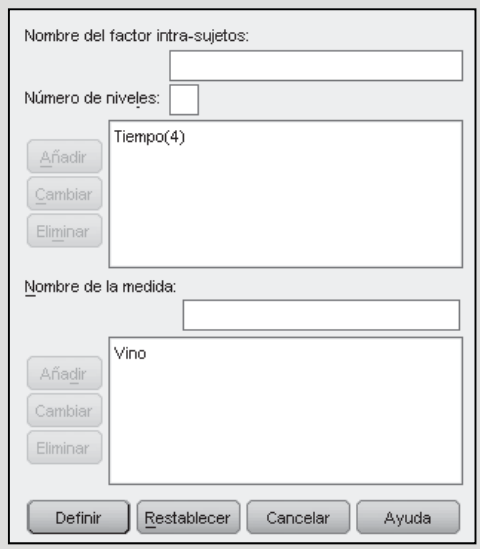
Paso 1. Para este primer ejemplo se va a estudiar si la capacidad de recordar la procedencia de un vino varía en el tiempo. Por lo tanto, tenemos un factor, que sería el tiempo, con cuatro niveles (una hora, un día, una semana y un mes) y la variable dependiente que es la capacidad de recordar la procedencia del vino. Se selecciona «Analizar» ⇨ «Modelo lineal general» ⇨ «Medidas repetidas…».


Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana Ayuda
Informes
Estadísticos descriptivos
Tablas
Comparar medias
Modelo lineal general
Modelos lineales generalizados
Modelos mixtos
Correlaciones
Regresión
Loglineal
Clasificar
Reducción de dimensiones
Escala
Pruebas no paramétricas
Predicciones
Superviv.
Respuesta múltiple
Control de calidad
Curva COR...
Univariante...
Multivariante...
Medidas repetidas...
Componentes de la varianza...
Individuo Edad A_Semana A_Mes V_Hora V_Día V_Semana
1 1,00 46,00 10,4 9,2 7,3
2 2,00 41,00 10,5 9,2 7,3
3 3,00 47,00 10,3 8,1 6,2
4 4,00 48,00 10,2 8,4 5,2
5 5,00 56,00 10,6 9,5 8,3
6 6,00 54,00 9,1 8,4 6,4 10,7 9,2 9,3
7 7,00 52,00 8,4 8,5 7,5 10,1 9,2 8,1
8 8,00 42,00 8,5 5,3 5,2 9,2 8,3 5,2
9 9,00 48,00 7,3 6,2 6,2 9,6 7,3 6,2
10 10,00 56,00 8,2 8,6 7,3 10,7 9,2 9,6
11 11,00 38,00 4,6 4,7 5,3 9,1 6,2 3,7
12 12,00 36,00 6,7 3,1 2,1 9,4 4,3 3,1
13 13,00 38,00 2,1 1,2 1,2 8,5 3,1 2,4
14 14,00 39,00 5,4 5,6 5,2 8,2 3,1 2,5
15 15,00 44,00 2,5 1,7 3,6 7,2 5,2 4,2

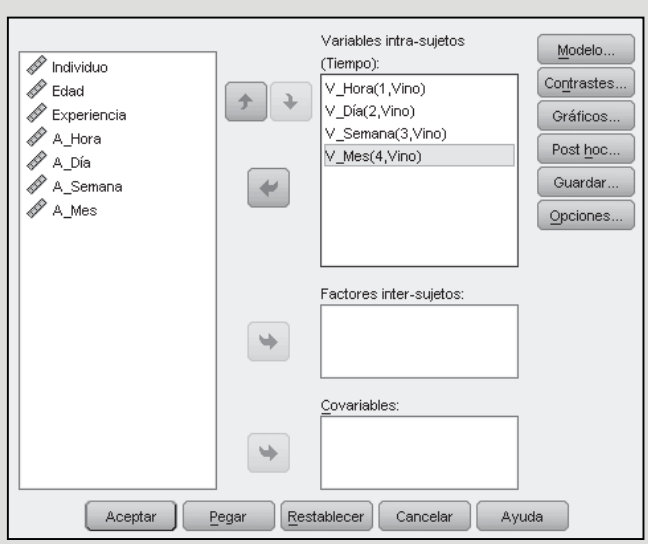
Paso 2. Es necesario definir los factores que tienen medidas repetidas. En nuestro ejemplo es el tiempo, por ello en «Nombre del factor intra-sujetos» se pone «Tiempo» (aunque se le podría poner cualquier nombre) y en «Número de niveles» se pone «4», ya que hay cuatro tiempos. En «Nombre de la medida» ponemos «Vino» para identificar la variable dependiente que se va a analizar. En ambos casos se pulsa «Añadir» para incorporar la información y, por último, pulsamos «Definir».

CUADRO IX.5. *(Continuación)*[SPSS]

Si hubiese otro factor con medidas repetidas también se añadiría en «Nombre del factor intra-sujetos» y, en este caso, se realizaría un ANOVA multifactorial. Si quisiéramos estudiar la capacidad de identificar la procedencia del aceite en el tiempo, además de la del vino, en «Nombre de la medida» se añadiría también «Aceite» y, en ese caso, estaríamos haciendo un ANOVA unifactorial con dos variables dependientes (multivariado).

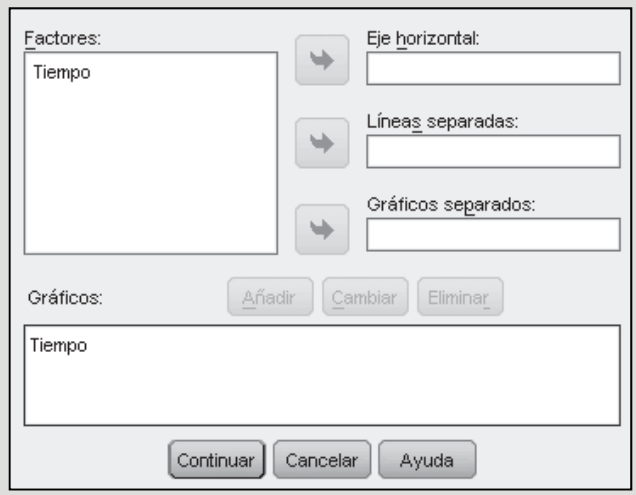


Paso 3. Ahora es necesario incorporar cada una de las variables dependientes a las «Variables intra-sujetos»: «V_Hora» a «_?_(1,Vino)», «V_Día» a «_?_(2,Vino)», «V_Semana» a «_?_(3,Vino)» y «V_Mes» a «_?_(4,Vino)».

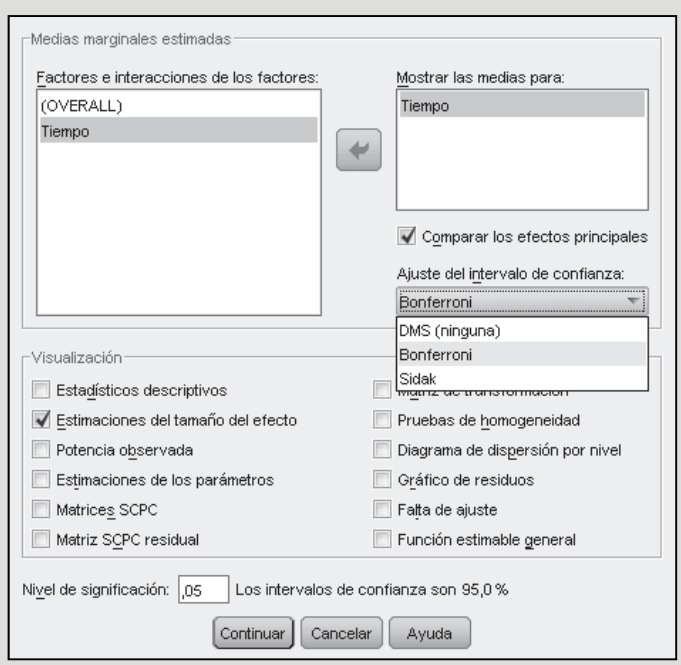


CUADRO IX.5. *(Continuación)*[SPSS]

Paso 4. En «Gráficos» se incluye el factor «Tiempo» en «Eje horizontal» y se pulsa «Añadir».



Paso 5. En «Opciones» se marca «Estimaciones del tamaño del efecto», que permite cuantificar la importancia de cada factor sobre la variable dependiente por medio del estadístico Eta al cuadrado parcial. La pestaña "post hoc" se utiliza solamente para comparar los niveles de los factores inter-sujetos. Para comparar dos a dos los niveles del factor intra-sujetos «Tiempo» se introduce en «Mostrar las medias para:», se marca «Comparar los efectos principales» y se selecciona «Bonferroni».



CUADRO IX.5. *(Continuación)*[SPSS]

Paso 6. Los resultados se pueden ver de dos formas diferentes. La primera es usando el estadístico *F* univariado. Un requisito de este estadístico en los ANOVAs con medidas repetidas es que las varianzas de las diferencias entre cada dos niveles del factor intra-sujetos, en este caso el tiempo, deben ser iguales, y los niveles independientes entre sí. Este supuesto equivale a afirmar que la matriz de varianzas-covarianzas es proporcional a la matriz unidad (varianzas iguales y covarianzas nulas), o que la dispersión conjunta es esférica. La siguiente tabla muestra los resultados de la prueba de esfericidad de Mauchley. Como la probabilidad es mayor que 0,05 (p = 0,073) se acepta la hipótesis de esfericidad. Si no se cumpliera el requisito de esfericidad (el valor p es menor que 0,05), es posible usar alguna modificación del test, Greenhouse-Geisser, Huynh-Feldt, o límite inferior, para corregir los grados de libertad a la baja lo que permite ajustar (incrementar) el valor p del contraste ANOVA.

Prueba de esfericidad de Mauchly[b]

Medida:Vino

Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Límite-inferior
dimension1 Tiempo	,547	10,084	5	,073	,700	,795	,333

El estadístico epsilon (G-G épsilon, 0,7; H-F épsilon, 0,795) expresa el grado en que la matriz de varianzas covarianzas se aleja de la esfericidad: vale 1 en condiciones de esfericidad perfecta, y cuanto más pequeño mayor es la desviación de la esfericidad. Una tercera corrección que también podemos utilizar se basa en el valor mínimo que puede tomar épsilon, que es igual al inverso del número de repeticiones menos uno (en este caso 1/(4-1) = 0,333).

Paso 7. Los resultados del estadístico *F* univariado se muestran a continuación. Tanto en el caso de esfericidad asumida, como en el caso de los test corregidos (para cuando no se cumple el requisito de esfericidad), se concluye que la capacidad de reconocer la procedencia del vino varía con el tiempo (p < 0,001).

Pruebas de efectos intra-sujetos.

Medida:Vino

Origen		Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial
Tiempo	Esfericidad asumida	245,243	3	81,748	77,633	,000	,812
	Greenhouse-Geisser	245,243	2,101	116,739	77,633	,000	,812
	Huynh-Feldt	245,243	2,385	102,840	77,633	,000	,812
	Límite-inferior	245,243	1,000	245,243	77,633	,000	,812
Error(Tiempo)	Esfericidad asumida	56,862	54	1,053			
	Greenhouse-Geisser	56,862	37,814	1,504			
	Huynh-Feldt	56,862	42,925	1,325			
	Límite-inferior	56,862	18,000	3,159			

CUADRO IX.5. *(Continuación)*[SPSS]

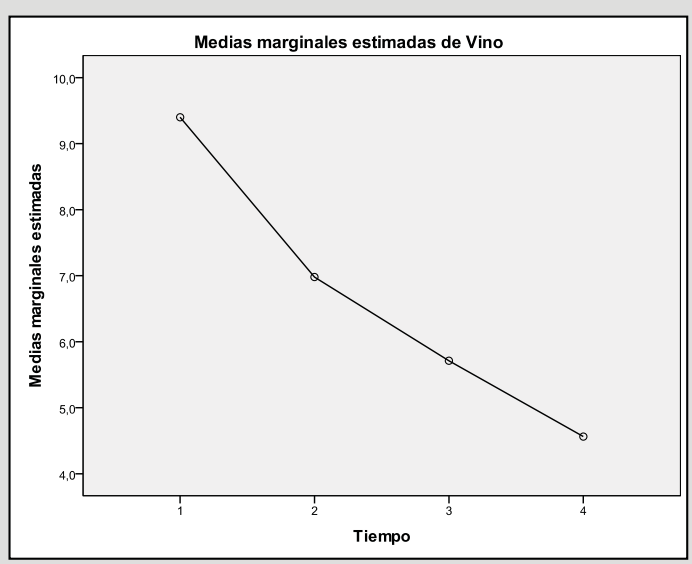
El estadístico Eta cuadrado parcial informa sobre el porcentaje de la variabilidad residual de la variable dependiente que es explicada por el factor tiempo, en un rango de 0 a 1. El tiempo tiene una gran influencia en la capacidad de recordar la procedencia del vino, ya que el valor es alto (0,812).

Paso 8. La segunda forma de ver los resultados es usando los estadísticos multivariados que se muestran a continuación. Cualquiera de los cuatro estadísticos no requiere la condición de esfericidad. Se observa que para los cuatros estadísticos la $p < 0{,}001$ y, por tanto, al igual que antes, se rechaza la hipótesis nula y se concluye que existen diferencias en el tiempo en la capacidad de reconocer la procedencia del vino. Es importante mencionar que, incluso en el supuesto de no cumplirse el requisito de esfericidad, si el tamaño de la muestra es pequeño, es mejor usar los estadísticos *F* univariados del paso 7, con cualquiera de los índices correctores épsilon, ya que la prueba es más potente.

Contrastes multivariados[b]

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Sig.	Eta al cuadrado parcial
Tiempo	Traza de Pillai	,883	40,123[a]	3,000	16,000	,000	,883
	Lambda de Wilks	,117	40,123[a]	3,000	16,000	,000	,883
	Traza de Hotelling	7,523	40,123[a]	3,000	16,000	,000	,883
	Raíz mayor de Roy	7,523	40,123[a]	3,000	16,000	,000	,883

Paso 9. El gráfico muestra claramente como el grado de acierto en la identificación de la procedencia del vino disminuye con el tiempo.



CUADRO IX.5. *(Continuación)*[SPSS]

Paso 10. La comparación dos a dos de los niveles del factor muestra que también existen diferencias significativas ($p < 0{,}001$), en todos los casos, entre los diferentes tiempos en el grado de acierto sobre la procedencia del vino.

Comparaciones por pares

Medida:Vino

(I)Tiempo	(J)Tiempo	Diferencia de medias (I-J)	Error típ.	Sig.[a]	Intervalo de confianza al 95 % para la diferencia[a]	
					Límite inferior	Límite superior
1	2	2,421*	,335	,000	1,428	3,414
	3	3,689*	,409	,000	2,478	4,900
	4	4,837*	,422	,000	3,588	6,086
2	1	-2,421*	,335	,000	-3,414	-1,428
	3	1,268*	,271	,001	,466	2,071
	4	2,416*	,288	,000	1,564	3,268
3	1	-3,689*	,409	,000	-4,900	-2,478
	2	-1,268*	,271	,001	-2,071	-,466
	4	1,147*	,228	,001	,472	1,822
4	1	-4,837*	,422	,000	-6,086	-3,588
	2	-2,416*	,288	,000	-3,268	-1,564
	3	-1,147*	,228	,001	-1,822	-,472

Basadas en las medias marginales estimadas.

*. La diferencia de medias es significativa al nivel ,05.

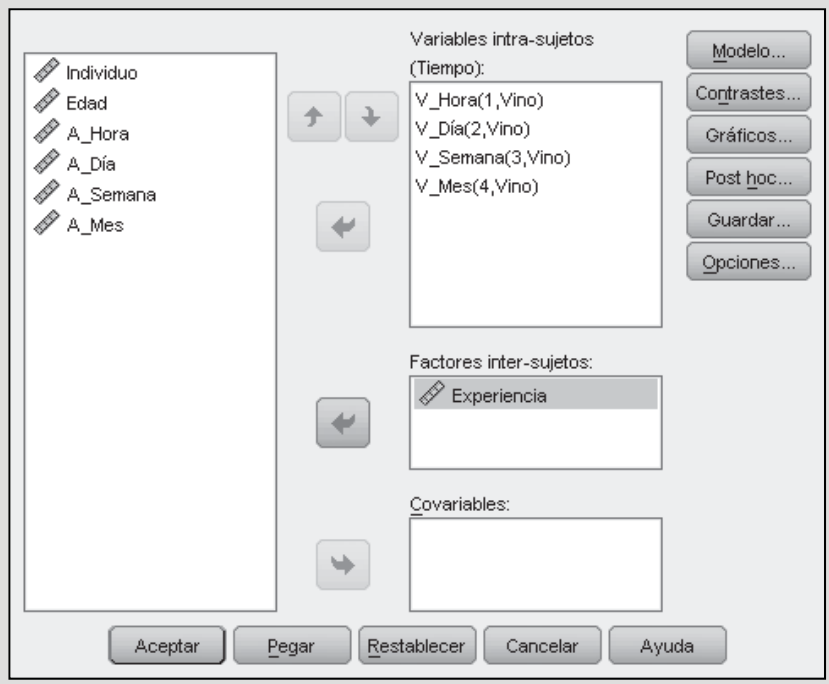
a. Ajuste para comparaciones múltiples: Bonferroni.

ANOVA multifactorial en el cual hay factores con medidas repetidas (factores intra-sujetos) y factores sin medidas repetidas (factores inter-sujetos)

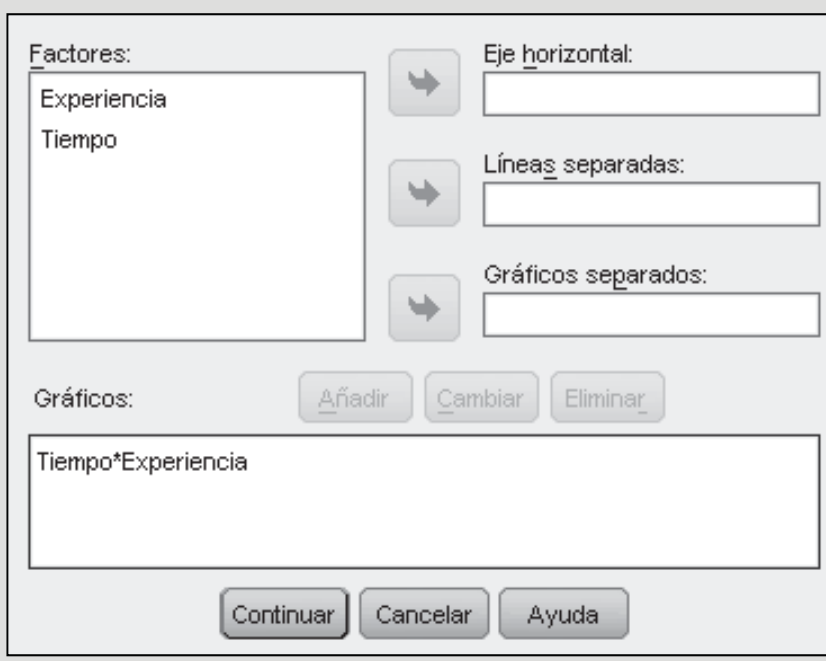
En este caso analizaremos de nuevo si la capacidad de recordar la procedencia de un vino varía en el tiempo (factor intra-sujeto), pero analizando si existen diferencias entre expertos y no expertos catadores (factor inter-sujetos).

Paso 11. Se siguen los pasos 1 a 3 para introducir el factor tiempo y luego se introduce el factor «Experiencia» en «Factores inter-sujetos», como se muestra en la siguiente ventana.

CUADRO IX.5. *(Continuación)*[SPSS]



Paso 12. En «Gráficos» se incluye el factor «Tiempo» en «Eje horizontal», «Experiencia» en «Líneas separadas» y se pulsa «Añadir».



CUADRO IX.5. *(Continuación)*[SPSS]

Paso 13. En «Opciones» se seleccionan los mismos estadísticos que se indicaron en el paso 5 y, además, hay que marcar la opción «Pruebas de homogeneidad». Como en este caso hay un factor inter-sujetos, se podrían hacer pruebas Post hoc, pero no tiene sentido en este caso ya que el factor «Experiencia» solo tiene dos niveles (expertos y no expertos).

Paso 14. Todos los estadísticos multivariados (véase paso 8 para más detalles) muestran diferencias significativas en el tiempo ($p < 0{,}001$) y existe interacción entre el tiempo y la experiencia ($p = 0{,}008$), lo cual significa que los cambios que se observan en el tiempo en la capacidad de acertar la procedencia del vino, dependen de la experiencia del catador.

Contrastes multivariados[b]

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Sig.	Eta al cuadrado parcial
Tiempo	Traza de Pillai	,918	56,267[a]	3,000	15,000	,000	,918
	Lambda de Wilks	,082	56,267[a]	3,000	15,000	,000	,918
	Traza de Hotelling	11,253	56,267[a]	3,000	15,000	,000	,918
	Raíz mayor de Roy	11,253	56,267[a]	3,000	15,000	,000	,918
Tiempo * Experiencia	Traza de Pillai	,534	5,724[a]	3,000	15,000	,008	,534
	Lambda de Wilks	,466	5,724[a]	3,000	15,000	,008	,534
	Traza de Hotelling	1,145	5,724[a]	3,000	15,000	,008	,534
	Raíz mayor de Roy	1,145	5,724[a]	3,000	15,000	,008	,534

Paso 15. Como se explicó en el paso 6 un requisito del estadístico *F* univariado es que las varianzas de las diferencias entre cada dos niveles del factor intrasujetos deben ser iguales. Como en la prueba de esfericidad de Mauchly la probabilidad es mayor que 0,05 ($p = 0{,}075$), se acepta la hipótesis de esfericidad.

Prueba de esfericidad de Mauchly[b]

Medida:Vino

Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Límite-inferior
Tiempo	,528	10,035	5	,075	,695	,840	,333

Paso 16. Los resultados del estadístico *F* univariado confirman los resultados de los estadísticos multivariados. Como se cumple el requisito de esfericidad, se utiliza el análisis en que se asume este requisito y se concluye que la capacidad de reconocer la procedencia del vino varía con el tiempo ($p < 0{,}001$). El estadístico Eta cuadrado parcial muestra que el tiempo tiene una gran influencia en la capacidad de recordar la procedencia del vino, ya que el valor es alto (0,848). Los resultados también corroboran la interacción entre tiempo y experiencia ya que $p < 0{,}05$ ($p = 0{,}006$).

CUADRO IX.5. *(Continuación)*[SPSS]

Pruebas de efectos intra-sujetos.

Medida:Vino

Origen		Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial
Tiempo	Esfericidad asumida	249,393	3	83,131	95,035	,000	,848
	Greenhouse-Geisser	249,393	2,084	119,664	95,035	,000	,848
	Huynh-Feldt	249,393	2,521	98,939	95,035	,000	,848
	Límite-inferior	249,393	1,000	249,393	95,035	,000	,848
Tiempo * Experiencia	Esfericidad asumida	12,250	3	4,083	4,668	,006	,215
	Greenhouse-Geisser	12,250	2,084	5,878	4,668	,015	,215
	Huynh-Feldt	12,250	2,521	4,860	4,668	,010	,215
	Límite-inferior	12,250	1,000	12,250	4,668	,045	,215
Error(Tiempo)	Esfericidad asumida	44,612	51	,875			
	Greenhouse-Geisser	44,612	35,430	1,259			
	Huynh-Feldt	44,612	42,852	1,041			
	Límite-inferior	44,612	17,000	2,624			

Paso 17. La siguiente tabla confirma el efecto del factor experiencia de las personas ya que existen diferencias significativas a la hora de reconocer la región de procedencia del vino entre los expertos y los no expertos ($F_{1,17} = 56{,}8$, $p < 0{,}001$) y el Eta cuadrado parcial muestra que el efecto de la experiencia es alto (0,77).

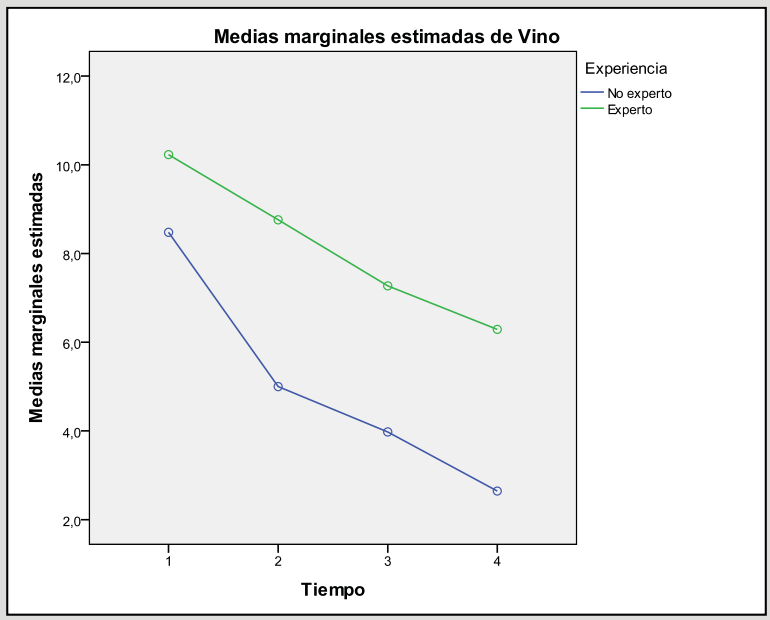
Pruebas de los efectos inter-sujetos

Medida:Vino
Variable transformada:Promedio

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial
Intersección	3282,658	1	3282,658	1016,557	,000	,984
Experiencia	183,556	1	183,556	56,843	,000	,770
Error	54,896	17	3,229			

Paso 18. La gráfica muestra claramente como el tiempo es importante ya que en ambos casos el grado de acierto al identificar la región del vino disminuye con el tiempo. En la gráfica también se observa el efecto significativo de la experiencia ya que esta disminución es menor en los expertos que en los no expertos.

CUADRO IX.5. *(Continuación)*[SPSS]



Paso 19. En concordancia con lo mostrado anteriormente, la comparación entre cada par de tiempos, del grado de acierto al identificar la región del vino, muestra diferencias significativas en todos los casos (p < 0,005).

Comparaciones por pares

Medida:Vino

(I)Tiempo	(J)Tiempo	Diferencia de medias (I-J)	Error típ.	Sig.[a]	Intervalo de confianza al 95 % para la diferencia[a]	
					Límite inferior	Límite superior
1	2	2,474*	,245	,000	1,743	3,205
	3	3,730*	,378	,000	2,603	4,857
	4	4,887*	,369	,000	3,786	5,987
2	1	-2,474*	,245	,000	-3,205	-1,743
	3	1,256*	,273	,002	,441	2,071
	4	2,413*	,296	,000	1,529	3,296
3	1	-3,730*	,378	,000	-4,857	-2,603
	2	-1,256*	,273	,002	-2,071	-,441
	4	1,157*	,231	,001	,468	1,845
4	1	-4,887*	,369	,000	-5,987	-3,786
	2	-2,413*	,296	,000	-3,296	-1,529
	3	-1,157*	,231	,001	-1,845	-,468

CUADRO IX.5. *(Continuación)*[SPSS]

Paso 20. El estadístico M de Box muestra que existe igualdad entre las matrices de varianzas-covarianzas (p = 0,056). El test de Levene también muestra que existe homogeneidad de varianzas entre todos los niveles, ya que la probabilidad es siempre mayor que 0,05. Por lo tanto, se cumple el requisito de la homogeneidad de varianzas entre los niveles del factor inter-sujeto.

Prueba de Box sobre la igualdad de las matrices de covarianzas[a]

M de Box	24,364
F	1,799
gl1	10
gl2	1339,524
Sig.	,056

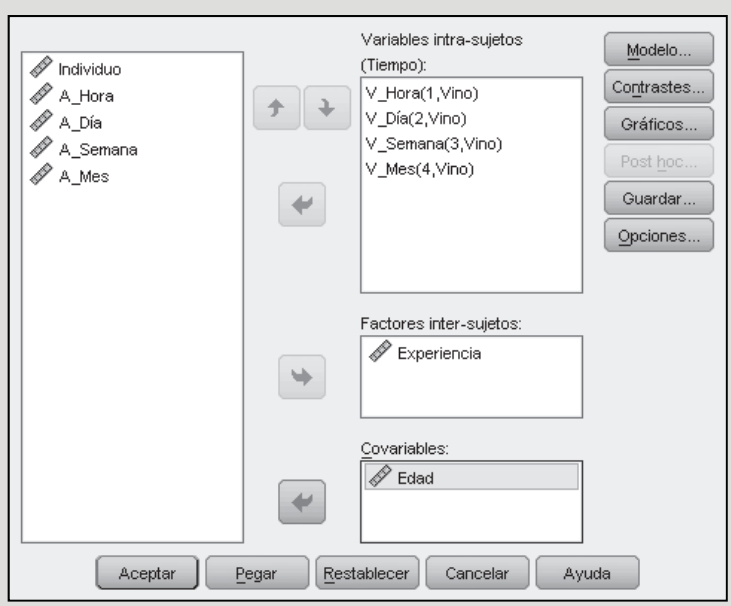
Contraste de Levene sobre la igualdad de las varianzas error[a]

	F	gl1	gl2	Sig.
V_Hora	,772	1	17	,392
V_Día	3,288	1	17	,109
V_Semana	,040	1	17	,844
V_Mes	,632	1	17	,438

Análisis de covarianza (ANCOVA)

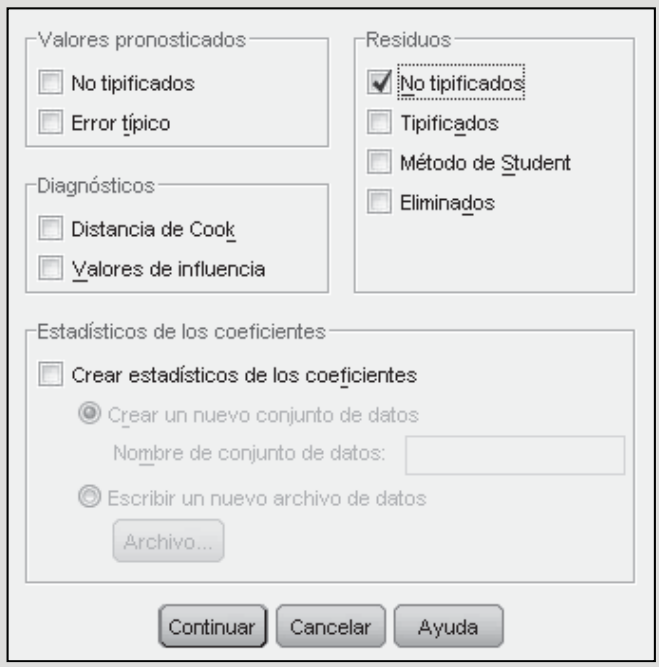
En el ejemplo anterior se observó que el grado de acierto al identificar la región de origen de un vino disminuía con el tiempo y que esta disminución era más acusada en las personas con menos experiencia en la cata de vinos. A continuación se quiere determinar si la edad de las personas puede ser un factor que influya en los resultados obtenidos con ambos factores, tiempo y experiencia.

Paso 21. Antes de realizar un ANCOVA es conveniente hacer primero el ANOVA, por tanto, seguimos con el ejemplo anterior, y simplemente se considera al factor «Edad» como una covariable y los otros factores se dejan como estaban.



CUADRO IX.5. *(Continuación)*[SPSS]

Paso 22. En la ventana anterior, en «Guardar», se marcan los residuos «No tipificados», para posteriormente ver si los residuos cumplen el requisito de Normalidad.



Paso 23. De nuevo todos los estadísticos multivariados (véase paso 8 para más detalles) muestran diferencias significativas en el tiempo (p < 0,001). Se sigue manteniendo el efecto significativo de la interacción entre el tiempo y la experiencia (p = 0,031), aunque es menor de lo mostrado anteriormente sin tener en cuenta el efecto de la edad. Por último, se observa un claro efecto de la edad (p = 0,005).

Efecto		Valor	F	Gl de la hipótesis	Gl del error	Sig.	Eta al cuadrado parcial
Tiempo	Traza de Pillai	,742	13,447[a]	3,000	14,000	,000	,742
	Lambda de Wilks	,258	13,447[a]	3,000	14,000	,000	,742
	Traza de Hotelling	2,882	13,447[a]	3,000	14,000	,000	,742
	Raíz mayor de Roy	2,882	13,447[a]	3,000	14,000	,000	,742
Tiempo * Edad	Traza de Pillai	,586	6,597[a]	3,000	14,000	,005	,586
	Lambda de Wilks	,414	6,597[a]	3,000	14,000	,005	,586
	Traza de Hotelling	1,414	6,597[a]	3,000	14,000	,005	,586
	Raíz mayor de Roy	1,414	6,597[a]	3,000	14,000	,005	,586
Tiempo * Experiencia	Traza de Pillai	,457	3,935[a]	3,000	14,000	,031	,457
	Lambda de Wilks	,543	3,935[a]	3,000	14,000	,031	,457
	Traza de Hotelling	,843	3,935[a]	3,000	14,000	,031	,457
	Raíz mayor de Roy	,843	3,935[a]	3,000	14,000	,031	,457

CUADRO IX.5. *(Continuación)*[SPSS]

Paso 24. Como en la prueba de esfericidad de Mauchly la probabilidad es mayor que 0,05 (p = 0,738), se acepta la hipótesis de esfericidad (véase paso 6 para más detalles).

Prueba de esfericidad de Mauchly[b]

Medida:Vino

Efecto intra-sujetos	W de Mauchly	Chi-cuadrado aprox.	gl	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Límite-inferior
Tiempo	,829	2,758	5	,738	,883	1,000	,333

Paso 25. Los resultados del estadístico F univariado confirman los resultados de los estadísticos multivariados. Como se cumple el requisito de esfericidad, se utiliza el análisis en que se asume este requisito y se concluye que la capacidad de reconocer la procedencia del vino varía con el tiempo ($F_{3,48}$ = 19,12, p < 0,001). Los resultados también corroboran la interacción entre tiempo y edad (p < 0,001) y la interacción tiempo y experiencia (p = 0,033).

Origen		Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial
Tiempo	Esfericidad asumida	34,212	3	11,404	19,128	,000	,545
	Greenhouse-Geisser	34,212	2,650	12,908	19,128	,000	,545
	Huynh-Feldt	34,212	3,000	11,404	19,128	,000	,545
	Límite-inferior	34,212	1,000	34,212	19,128	,000	,545
Tiempo * Edad	Esfericidad asumida	15,995	3	5,332	8,943	,000	,359
	Greenhouse-Geisser	15,995	2,650	6,035	8,943	,000	,359
	Huynh-Feldt	15,995	3,000	5,332	8,943	,000	,359
	Límite-inferior	15,995	1,000	15,995	8,943	,009	,359
Tiempo * Experiencia	Esfericidad asumida	5,651	3	1,884	3,159	,033	,165
	Greenhouse-Geisser	5,651	2,650	2,132	3,159	,040	,165
	Huynh-Feldt	5,651	3,000	1,884	3,159	,033	,165
	Límite-inferior	5,651	1,000	5,651	3,159	,095	,165
Error(Tiempo)	Esfericidad asumida	28,617	48	,596			
	Greenhouse-Geisser	28,617	42,406	,675			
	Huynh-Feldt	28,617	48,000	,596			
	Límite-inferior	28,617	16,000	1,789			

CUADRO IX.5. *(Continuación)*[SPSS]

Si se produce incongruencia entre los resultados obtenidos en los estadísticos multivariados y univariados, en el caso de cumplirse el requisito de esfericidad debemos utilizar la aproximación univariada, ya que es más potente, sobre todo si el tamaño de la muestra es pequeño.

Paso 26. Aunque tanto el efecto de la edad ($F_{1,16}$ = 23,1, p < 0,001) como de la experiencia ($F_{1,16}$ = 77,28, p < 0,001) son significativos, el Eta cuadrado parcial es mayor para la experiencia (0,828) que para la edad (0,591).

Pruebas de los efectos inter-sujetos

Medida:Vino
Variable transformada:Promedio

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial
Intersección	2,352	1	2,352	1,676	,214	,095
Edad	32,439	1	32,439	23,112	,000	,591
Experiencia	108,470	1	108,470	77,282	,000	,828
Error	22,457	16	1,404			

Paso 27. La comparación entre cada par de tiempos, en el grado de acierto al identificar la región del vino, sigue mostrando diferencias significativas en todos los casos (p < 0,002).

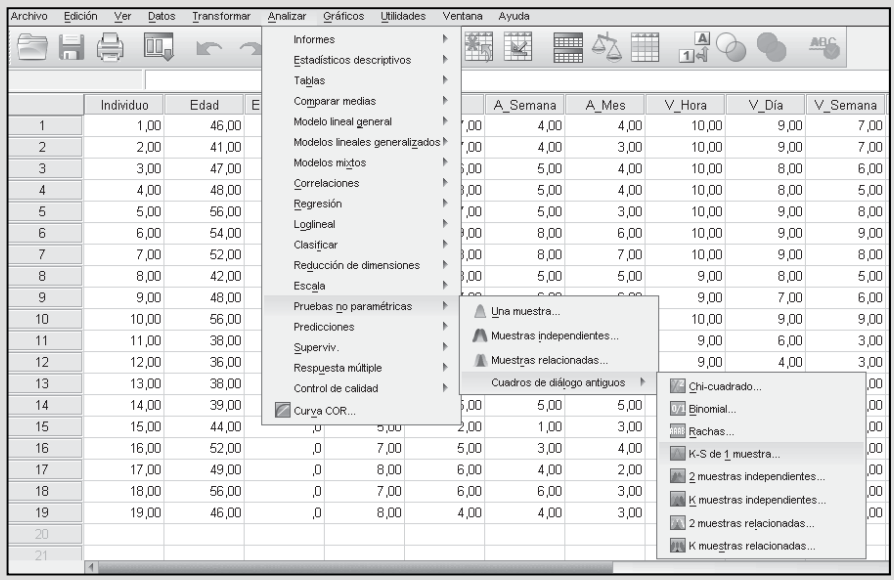
Comparaciones por pares

Medida:Vino

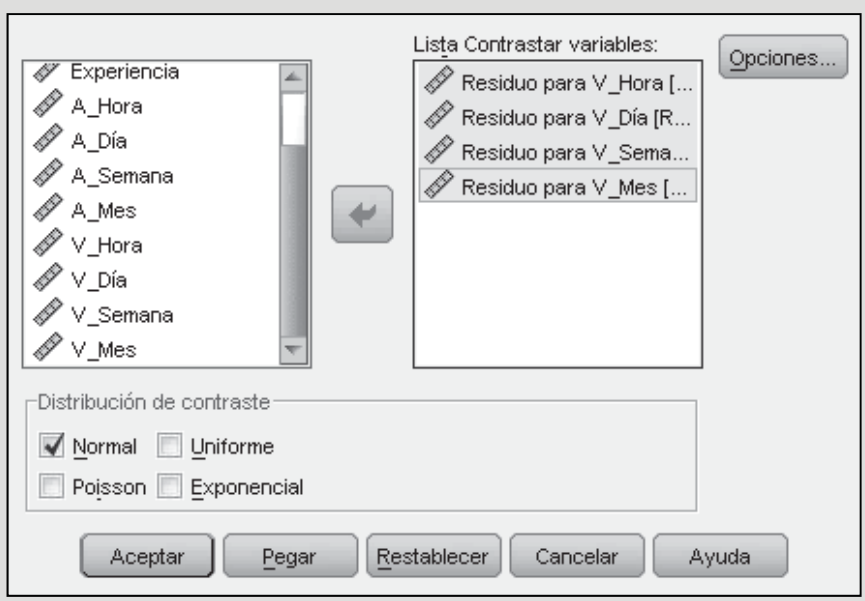
(I)Tiempo	(J)Tiempo	Diferencia de medias (I-J)	Error típ.	Sig.[a]	Intervalo de confianza al 95 % para la diferencia[a]	
					Límite inferior	Límite superior
1	2	2,463*	,217	,000	1,811	3,115
	3	3,705*	,252	,000	2,947	4,463
	4	4,865*	,287	,000	4,001	5,729
2	1	-2,463*	,217	,000	-3,115	-1,811
	3	1,242*	,227	,000	,559	1,925
	4	2,403*	,281	,000	1,557	3,248
3	1	-3,705*	,252	,000	-4,463	-2,947
	2	-1,242*	,227	,000	-1,925	-,559
	4	1,161*	,233	,001	,459	1,862
4	1	-4,865*	,287	,000	-5,729	-4,001
	2	-2,403*	,281	,000	-3,248	-1,557
	3	-1,161*	,233	,001	-1,862	-,459

CUADRO IX.5. *(Continuación)*[SPSS]

Paso 28. Para despejar cualquier duda con respecto al requisito de que los datos tengan una distribución Normal, analizamos la distribución de los residuos. Se selecciona «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «K-S de 1 muestra…».



Paso 29. En «Lista Contrastar variables:» se introducen los residuos y en «Distribución de contraste» se marca «Normal».



CUADRO IX.5. *(Continuación)*[SPSS]

Paso 30. Se observa que para los 4 tiempos, los residuos tienen una distribución Normal, ya que $p > 0,05$: $p = 0,925$ para «V_Hora», $p = 0,526$ para «V_Día», $p = 0,958$ para «V_Semana» y $p = 0,634$ para «V_Mes».

Prueba de Kolmogorov-Smirnov para una muestra

		Residuo para V_Hora	Residuo para V_Día	Residuo para V_Semana	Residuo para V_Mes
N		19	19	19	19
Parámetros normales[a,b]	Media	,0000	,0000	,0000	,0000
	Desviación típica	,56902	,83476	,97860	,92692
Diferencias más extremas	Absoluta	,126	,186	,117	,171
	Positiva	,073	,186	,091	,162
	Negativa	-,126	-,162	-,117	-,171
Z de Kolmogorov-Smirnov		,548	,811	,509	,746
Sig. asintót. (bilateral)		,925	,526	,958	,634

a. La distribución de contraste es la Normal.
b. Se han calculado a partir de los datos.

Paso 31. Por último, se cumple el requisito de homogeneidad de varianzas entre los niveles de los factores inter-sujeto (M de Box, $p = 0,056$; test de Levene, $p > 0,05$ en todos los casos). Ambas pruebas contrastan la homogeneidad entre niveles del factor inter-sujetos, la M de Box todas las variables al mismo tiempo, el test de Levene cada variable por separado; la homogeneidad intra-sujetos es la esfericidad ya comprobada anteriormente, que también se cumple.

Prueba de Box sobre la igualdad de las matrices de covarianzas[a]

M de Box	24,364
F	1,799
gl1	10
gl2	1339,524
Sig.	,056

Contraste de Levene sobre la igualdad de las varianzas error[a]

	F	gl1	gl2	Sig.
V_Hora	1,283	1	17	,466
V_Día	,273	1	17	,395
V_Semana	,410	1	17	,366
V_Mes	1,779	1	17	,244

IX.1.7. ANÁLISIS DE VARIANZA ANIDADO

El ANOVA anidado es un diseño experimental en el que los niveles de uno o más factores dependen de los niveles de otros factores. Es una extensión de un ANOVA de un factor, en el cual se divide cada grupo en subgrupos. Se dice que un factor B está anidado en otro factor A (o que sus niveles están anidados en los de A) cuando todos los niveles del factor B aparecen asociados a un único nivel del factor A. Se denota como $B \subset A$. Es posible tener varios factores anidados en el diseño experimental, es decir, tener grupos, subgrupos para cada grupo, subsubgrupos para cada subgrupo, etc.

El análisis de varianza anidado tiene una hipótesis nula para cada factor. En un ANOVA anidado de dos vías, una hipótesis nula sería que los subgrupos dentro de cada grupo tienen la misma media, la segunda hipótesis nula sería que los grupos tienen la misma media. Los requisitos del ANOVA anidado son iguales a los del ANOVA, es decir, se requiere que los residuos tengan distribución Normal y homogeneidad de varianzas.

Un ejemplo de un ANOVA anidado es plantar una especie de árbol en tres hábitats diferentes, por ejemplo, en el valle, en la ladera de la montaña y en la alta montaña, y repetir ese diseño experimental en dos zonas diferentes, por ejemplo, una zona que es un parque natural y, por tanto, existe cierto grado de protección, y otra zona que no tiene ningún tipo de control. El objetivo sería determinar si existen variaciones en la tasa de crecimiento de los árboles (variable dependiente) entre las dos zonas (factor primario), y entre los hábitats dentro de cada zona (factor secundario anidado).

En otro ejemplo, supongamos que se está probando la hipótesis nula de si monos sometidos a estrés y no sometidos a estrés tienen la misma cantidad de anticuerpos. Como los monos ocupan mucho espacio, es necesario tenerlos en cuatro habitaciones diferentes y en cada una de ellas hay monos de los dos tipos, a los que se les ha sometido y no ha sometido a estrés. Se puede dar el caso de que las condiciones de las habitaciones (luz, temperatura, etc.) pueden causar algunas diferencias en el nivel de estrés y, por tanto, tal vez en el nivel de anticuerpos. Para eliminar este potencial efecto de las diferentes habitaciones en el diseño experimental, se debe hacer una ANOVA anidado en el que el primer factor sería las habitaciones (con cuatro niveles, las 4 habitaciones) y el factor anidado sería el nivel de estrés de los monos (con dos niveles, estresados y no estresados). Con este diseño se puede ver si existen diferencias en la cantidad de anticuerpos (variable dependiente) entre habitaciones y en la cantidad de anticuerpos entre el grupo de monos estresados y no estresados dentro de cada habitación.

Además de informar si existen diferencias en las medias entre los niveles de cada factor, en las medias entre cada nivel del subfactor dentro de cada factor, etc., el ANOVA anidado es muy útil para futuros diseños experimentales, ya que informa del grado de variabilidad dentro de cada factor. Por ejemplo, imaginemos que se miden los niveles de glucógeno en varios tipos de músculos en ratas sometidas y no sometidas a estrés. Por tanto, el factor músculo estaría anidado dentro del factor estrés y la variable dependiente sería el nivel de glucógeno. Se puede dar el caso de que la variabilidad en el contenido de glucógeno entre músculos sea muy pequeña, mientras que entre ratas sea alta. Por tanto, en el siguiente diseño habría que centrar el esfuerzo en medir el glucógeno en un mayor número de ratas y tan solo en un músculo dentro de cada rata, es decir, ya no sería necesario un ANOVA anidado pues nos quedaríamos con un solo factor, el nivel de estrés al que se somete a las ratas.

CUADRO IX.6. ANOVA anidado

Aplicación con R.

EJEMPLO. Los líquenes son organismos considerados bioindicadores de contaminación, ya que se ha demostrado que su presencia suele ser menor en zonas más contaminadas. En el archivo **Cuadro IX.6.V.csv** hay datos del promedio de especies de líquenes encontrados en dos especies de árbol, *Populus alba* (Chopo blanco) y *Aesculus hippocastanum* (Castaño de indias), en tres ciudades de España (Madrid, Barcelona y Sevilla) y en tres tipos de ambientes (poca, media y alta contaminación). El promedio es el resultado de contar las especies de líquenes en 40 árboles de cada una de las especies (*Populus alba* y *Aesculus hippocastanum*), en cada una de las ciudades y en los tres tipos de ambientes. El objetivo es determinar si el número de especies de líquenes varía en función del grado de contaminación. Como el número de especies de líquenes puede verse afectado por la especie de árbol y/o por el grado de contaminación de las ciudades, es necesario realizar un ANOVA anidado.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro IX.6.R**, que se muestra en la siguiente ventana. Sólo es necesario especificar lo que se solicita en la sección modificable por el usuario. Para que este *script* funcione correctamente es necesario tener instalados los paquetes "e1071", "class", "car", "lawstat", "mvtnorm", "VGAM", "splines", "stats4", "nortest", "multcomp" y "survival", o realizar la instalación como se explica en el Apéndice II.

```
#Cuadro IX.6
#Análisis de varianza anidado.

################################################
#Sección modificable por el usuario
################################################

#Lectura de datos.
datos<-read.csv2("Cuadro IX.6.V.csv",header=TRUE,encoding="latin1")

#Selección variable dependiente. En caso de que se deseen utilizar valo
varInteres<-c("Número.de.especies.de.líquenes")

#Selección de un tipo de transformación para las variables de interés.
varInteres1X2<-NULL
varInteres1X<-NULL
varInteresLN<-NULL
varInteresLOG10<-NULL
varInteresSQR<-NULL
varInteresX2<-NULL
varInteresX3<-NULL
varInteresEX<-NULL
varInteresASN<-NULL

#varInteres1X2<-c(1,1)
#varInteres1X<-c(1,1)
#varInteresLN<-c(1,1)
#varInteresLOG10<-c(1,1)
#varInteresSQR<-c(1,1)
#varInteresX2<-c(1,1)
#varInteresX3<-c(1,1)
#varInteresEX<-c(1,1)
#varInteresASN<-c(1,1)
```

CUADRO IX.6. *(Continuación)*[R]

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro IX.6.V.csv**.

```
datos<-read.csv2("Cuadro IX.6.V.csv",header=TRUE,encoding="latin1")
```

Paso 2. Se selecciona la variable dependiente, en este caso el número de líquenes.

```
varInteres<-c("Número.de.especies.de.líquenes")
```

Paso 3. Cuando sea necesario es posible transformar la variable dependiente, para ello simplemente hay que quitar el símbolo # en el tipo de transformación que se quiera realizar. En primer lugar se hace el análisis sin transformar los datos, y en el caso de no cumplirse la Normalidad y/o homogeneidad de varianzas, entonces se realizaría el análisis aplicando la transformación más adecuada.

```
#varInteres1X2<-c(1,1)
#varInteres1X<-c(1,1)
#varInteresLN<-c(1,1)
#varInteresLOG10<-c(1,1)
#varInteresSQR<-c(1,1)
#varInteresX2<-c(1,1)
#varInteresX3<-c(1,1)
#varInteresEX<-c(1,1)
#varInteresASN<-c(1,1)
```

Paso 4. A continuación se indica cuales son los factores con los que se va a trabajar. Es posible hacer el análisis con un máximo de 4 factores con este *script*. Como en este ejemplo solo trabajamos con tres factores, el factor 4 tiene activada la opción «NULL» ya que no tiene el símbolo # delante de la instrucción.

```
Factor1<-c("Ciudad")
Factor2<-c("Especie.Arbol")
Factor3<-c("Grado.Contaminación")
#Factor3<-NULL
#Factor4<-c("")
Factor4<-NULL
```

Paso 5. Lo más importante es definir correctamente el anidamiento de los factores. Si por ejemplo, solo hubiese un factor anidado en otro, en el paso anterior el factor anidado debería ser el factor 2 y el 1 el factor en que está anidado. En ese caso se escogería la opción 1 y se podrían poner los factores 3 y 4 como factores independientes no anidados. En nuestro caso hay 3 factores y dos de ellos están anidados, por ello se escoge la opción 3.

```
##Posibilidades de anidamiento de los factores
#Para 2 factores
        #1.Factor2 en Factor1
```

CUADRO IX.6. *(Continuación)*[R]

```
#Para 3 factores
        #2.Factor3 en Factor2
        #3.Factor3 en Factor2.Factor2 en Factor1
#Para 4 factores
        #4.Factor4 en Factor3
        #5.Factor4 en Factor3. Factor2 en Factor1
        #6.Factor4 en Factor3. Factor3 en Factor2
        #7.Factor4 en Factor3. Factor3 en Factor2. Factor2 en Factor1.

anidamiento<-3
```

Paso 6. También es posible seleccionar grupos en dos variables diferentes (un máximo de dos grupos en cada variable). En este ejemplo no se hace ninguna selección.

```
#Variables de selección.
#varSeleccion1<-list(c("","",""))
varSeleccion1<-NULL
#varSeleccion2<-list(c("","",""))
varSeleccion2<-NULL
```

Paso 7. Se suelen utilizar cuatro tipos de sumas de cuadrados: tipo I para modelos anidados, tipo II para modelos equilibrados, tipo III para modelos no equilibrados (o también equilibrados), y tipo IV para modelos con casillas vacías. Cuando el diseño es totalmente anidado puede ser adecuado utilizar sumas de cuadrados tipo I, pero no es incorrecto utilizar el tipo III. El tipo III es adecuado para la mayoría de las aplicaciones.

```
#SS<-c(2)
SS<-c(3)
```

Paso 8. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
ArchivodeSalida<-"Salida Cuadro IX.6.txt"
```

Paso 9. En primer lugar nos fijamos en la Normalidad de los residuos. Se observa que tanto el test Kolmogorov–Smirnov con la corrección de Lilliefors (p = 0,411), como el de Shapiro-Wilk (p = 0,4003) muestran p > 0,05 y, por tanto, los residuos se ajustan a una distribución Normal.

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos
D = 0.0492, p-value = 0.4211


[[5]]

        Shapiro-Wilk normality test

data:  residuos
W = 0.9912, p-value = 0.4003
```

CUADRO IX.6. *(Continuación)*[R]

El test de Levene considerando la mediana muestra que también se cumple la homogeneidad de varianzas (p = 0,098). Por lo tanto, no es necesario transformar los datos.

```
[[6]]
[[6]][[1]]
[1] "Factor1"

[[6]][[2]]

        modified robust Brown-Forsythe Levene-type test
        deviations from the median

data:  datos3$residuos
Test Statistic = 2.8556, p-value = 0.0604


[[6]][[3]]
[1] "Factor2"

[[6]][[4]]

        modified robust Brown-Forsythe Levene-type test
        deviations from the median

data:  datos3$residuos
Test Statistic = 1.9585, p-value = 0.1636


[[6]][[5]]
[1] "Factor3"

[[6]][[6]]

        modified robust Brown-Forsythe Levene-type test
        deviations from the median

data:  datos3$residuos
Test Statistic = 1.9874, p-value = 0.1404
```

Paso 10. Los resultados del ANOVA se muestran en la siguiente tabla. Existen diferencias significativas entre las ciudades en el número de especies de líquenes (ANOVA, $F_{2,\ 156}$ = 12,651, p < 0,001), pero no existen diferencias entre las dos especies de árboles (ANOVA, $F_{2,\ 156}$ = 1,46, p = 0,226). En lo que se refiere al objetivo del ejemplo, existen diferencias significativas en el número de especies de líquenes observadas en función del grado de contaminación (ANOVA, $F_{4,\ 156}$ = 94,347, p < 0,001).

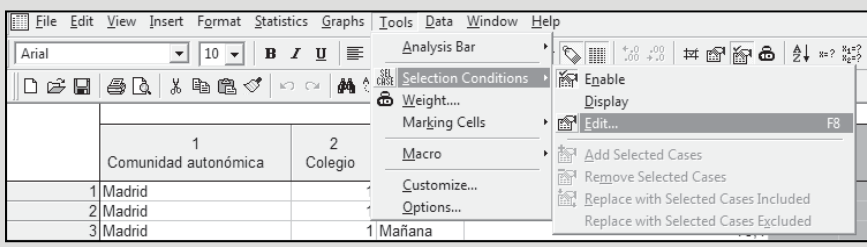
```
[[2]]
Anova Table (Type III tests)

Response: valores
             Sum Sq  Df   F value    Pr(>F)
(Intercept) 8046.1   1 2866.8073 < 2.2e-16 ***
F1            71.0   2   12.6509 8.098e-06 ***
F1:F2         12.3   3    1.4638    0.2266
F2:F3       1059.2   4   94.3472 < 2.2e-16 ***
Residuals    437.8 156
```

CUADRO IX.6. *(Continuación)*[STATISTICA]

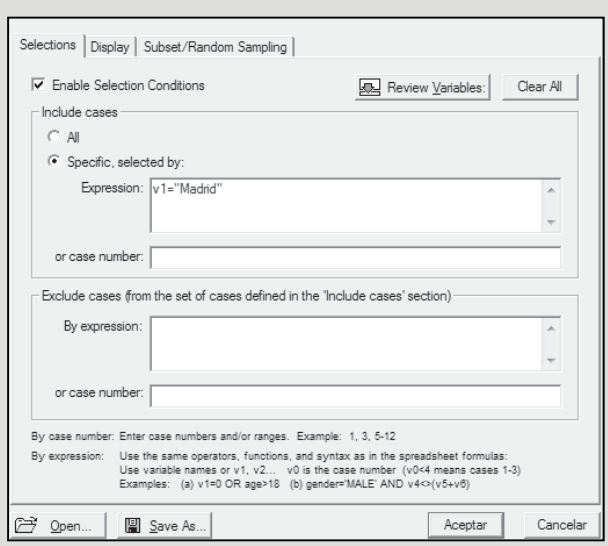
Aplicación con STATISTICA.

EJEMPLO. En el archivo **Cuadro IX.6.sta** hay datos del porcentaje de alumnos que aprueban todas las asignaturas en el año en clases de los turnos de mañana y tarde, en diferentes colegios de distintas comunidades autónomas de España. El objetivo es determinar si existen diferencias en el número de aprobados entre los turnos de mañana y tarde. El turno de tarde son estudiantes que normalmente trabajan, mientras que el turno de mañana son en su mayoría estudiantes que no trabajan. Como puede existir un efecto sobre el porcentaje de aprobados asociado al colegio y/o la comunidad autónoma, es necesario realizar un ANOVA anidado. Vamos a realizar el ejemplo considerando solo la Comunidad de Madrid y determinar el posible efecto del colegio, para cuantificar realmente si existen diferencias entre ambos turnos.

Paso 1. Debido a que existen varias comunidades autónomas, es necesario primero definir con la que se va a trabajar. Para seleccionar los casos vamos a «Tools» ⇨ «Selection Conditions» ⇨ «Edit…».

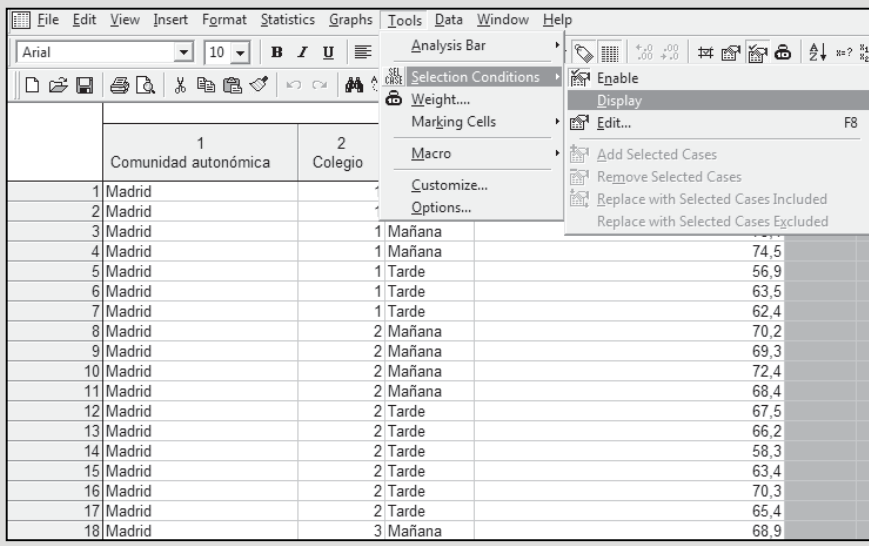


Paso 2. En la pestaña «Selections» se marca «Enable Selection Conditions» ⇨ «Specific, selected by:» donde se selecciona la variable o variables a analizar, usando OR y AND si se quieren combinar varias condiciones. En este caso se selecciona «v1="Madrid"».


Selections Display Subset/Random Sampling
Enable Selection Conditions
Review Variables:
Clear All
Include cases
All
Specific, selected by:
Expression: v1="Madrid"
or case number:
Exclude cases (from the set of cases defined in the 'Include cases' section)
By expression:
or case number:
By case number: Enter case numbers and/or ranges. Example: 1, 3, 5-12
By expression: Use the same operators, functions, and syntax as in the spreadsheet formulas:
Use variable names or v1, v2... v0 is the case number (v0<4 means cases 1-3)
Examples: (a) v1=0 OR age>18 (b) gender='MALE' AND v4<>(v5+v6)
Open...
Save As...
Aceptar
Cancelar

CUADRO IX.6. *(Continuación)*[STATISTICA]

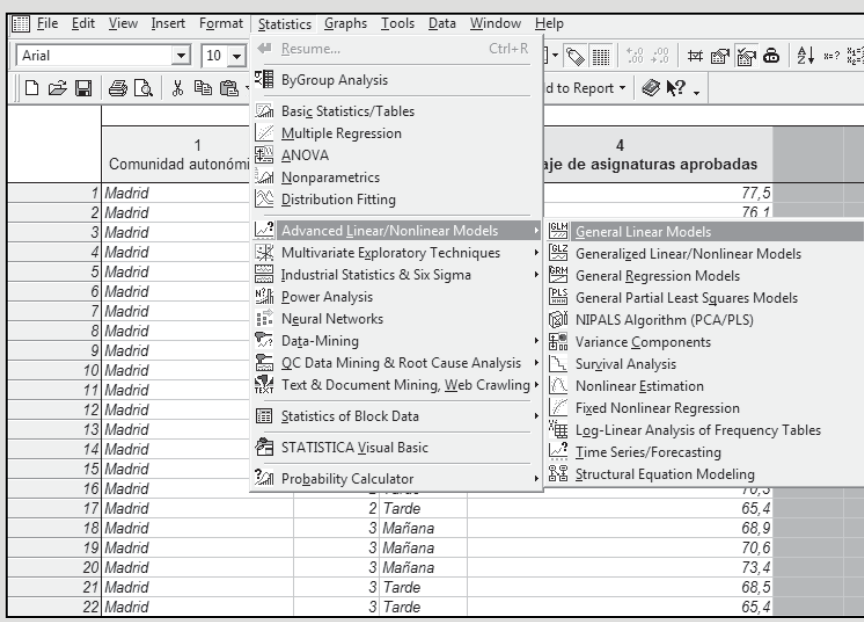
Paso 3. Para resaltar los casos seleccionados vamos a «Tools» ⇨ «Selection Conditions» ⇨ «Display». En la tabla de datos aparecerán en cursiva los valores seleccionados. En el supuesto de que no aparecieran en cursiva, eso significaría que la selección en el paso 2 no se hizo de forma correcta.



	1 Comunidad autonómica	2 Colegio	
1	Madrid		
2	Madrid		
3	Madrid	1	Mañana
4	Madrid	1	Mañana
5	Madrid	1	Tarde
6	Madrid	1	Tarde
7	Madrid	1	Tarde
8	Madrid	2	Mañana
9	Madrid	2	Mañana
10	Madrid	2	Mañana
11	Madrid	2	Mañana
12	Madrid	2	Tarde
13	Madrid	2	Tarde
14	Madrid	2	Tarde
15	Madrid	2	Tarde
16	Madrid	2	Tarde
17	Madrid	2	Tarde
18	Madrid	3	Mañana

	4
1	74,5
2	56,9
3	63,5
4	62,4
5	70,2
6	69,3
7	72,4
8	68,4
9	67,5
10	66,2
11	58,3
12	63,4
13	70,3
14	65,4
15	68,9

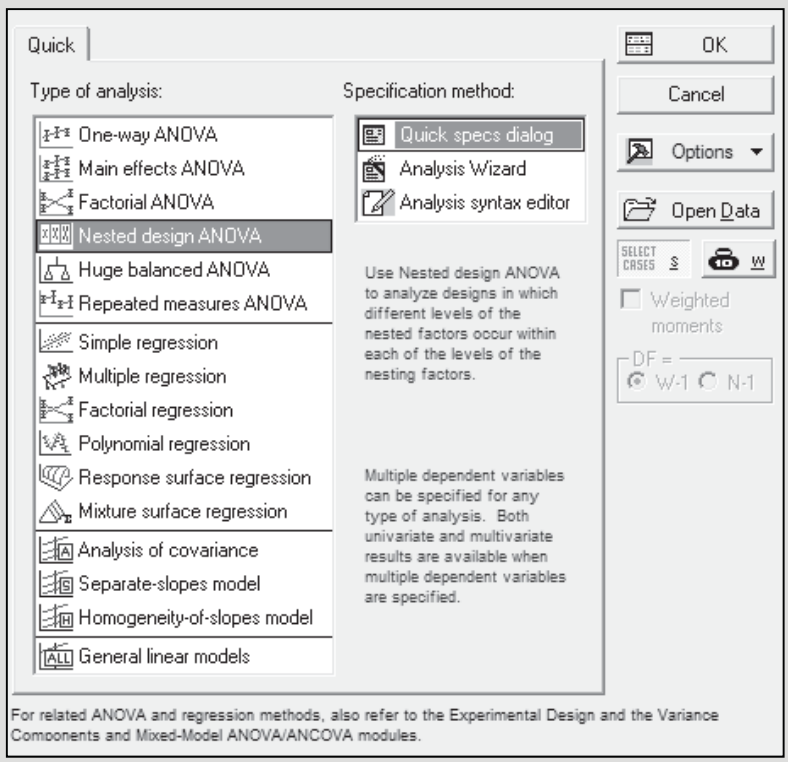
Paso 4. Para realizar el ANOVA se selecciona «Statistics» ⇨ «Advanced Linear/Nonlinear Models» ⇨ «General Linear Models».



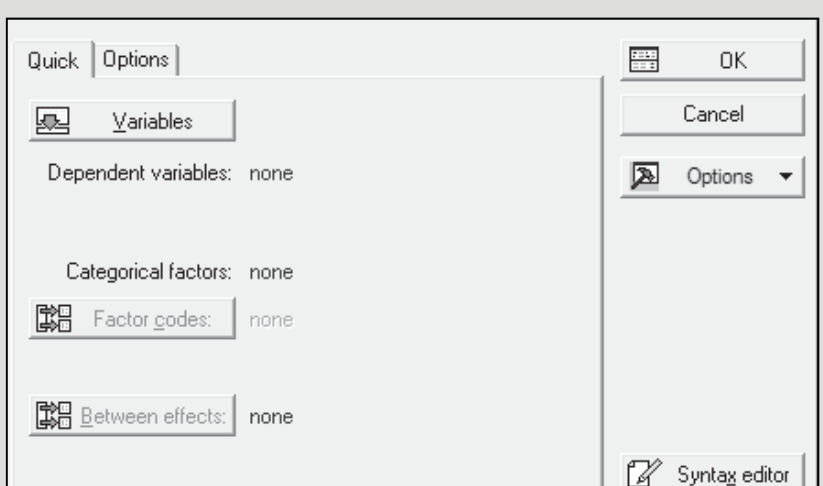
	1 Comunidad autonómi		4 aje de asignaturas aprobadas
1	Madrid		77,5
2	Madrid		76,1
3	Madrid		
4	Madrid		
5	Madrid		
6	Madrid		
7	Madrid		
8	Madrid		
9	Madrid		
10	Madrid		
11	Madrid		
12	Madrid		
13	Madrid		
14	Madrid		
15	Madrid		
16	Madrid		
17	Madrid	2 Tarde	65,4
18	Madrid	3 Mañana	68,9
19	Madrid	3 Mañana	70,6
20	Madrid	3 Mañana	73,4
21	Madrid	3 Tarde	68,5
22	Madrid	3 Tarde	65,4

CUADRO IX.6. *(Continuación)*[STATISTICA]

Paso 5. En «Type of analysis» se selecciona «Nested design ANOVA» y en «Specification method:» marcamos «Quick specs dialog».

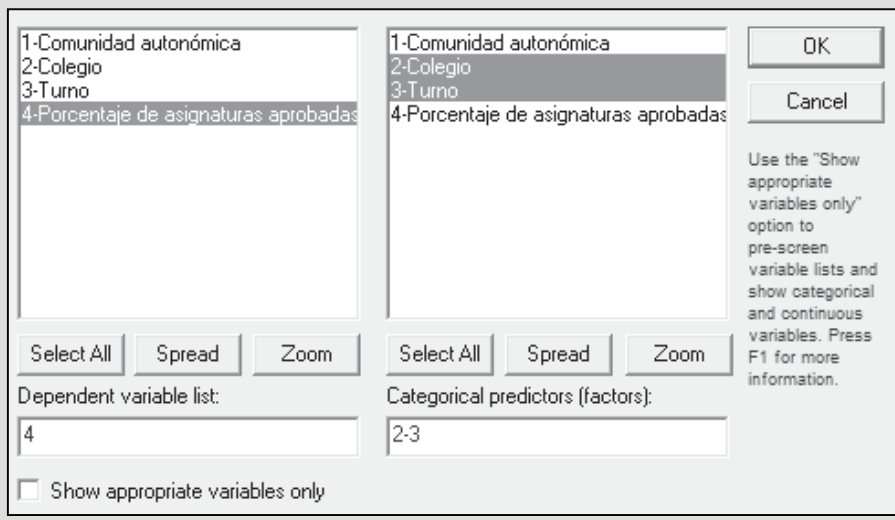

Quick
Type of analysis:
One-way ANOVA
Main effects ANOVA
Factorial ANOVA
Nested design ANOVA
Huge balanced ANOVA
Repeated measures ANOVA
Simple regression
Multiple regression
Factorial regression
Polynomial regression
Response surface regression
Mixture surface regression
Analysis of covariance
Separate-slopes model
Homogeneity-of-slopes model
General linear models
Specification method:
Quick specs dialog
Analysis Wizard
Analysis syntax editor
Use Nested design ANOVA to analyze designs in which different levels of the nested factors occur within each of the levels of the nesting factors.
Multiple dependent variables can be specified for any type of analysis. Both univariate and multivariate results are available when multiple dependent variables are specified.
OK
Cancel
Options
Open Data
SELECT CASES
Weighted moments
DF =
W-1
N-1
For related ANOVA and regression methods, also refer to the Experimental Design and the Variance Components and Mixed-Model ANOVA/ANCOVA modules.

Paso 6. En la siguiente ventana, en la pestaña «Quick» se pulsa en «Variables».

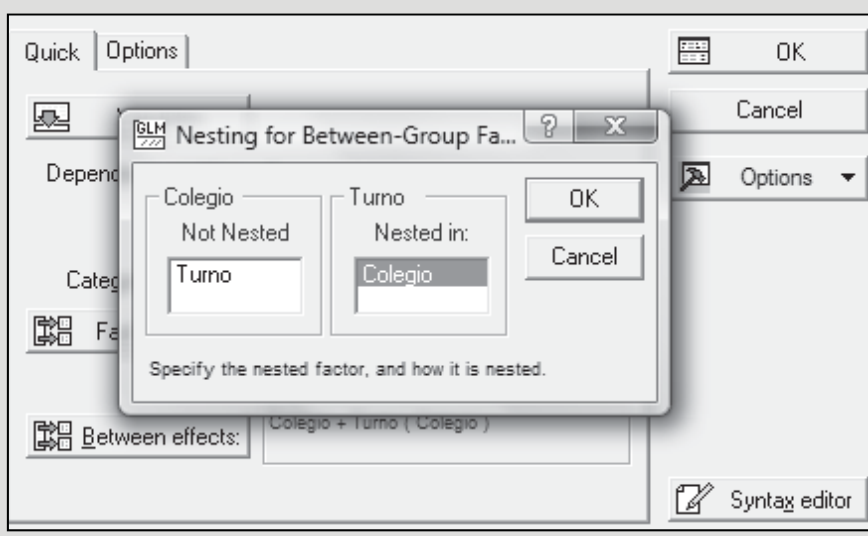


CUADRO IX.6. *(Continuación)*[STATISTICA]

Paso 7. En «Dependent variable list» introducimos «Porcentaje de asignaturas aprobadas» y en «Categorical predictor (factors):» introducimos los factores «Colegio» y «Turno».



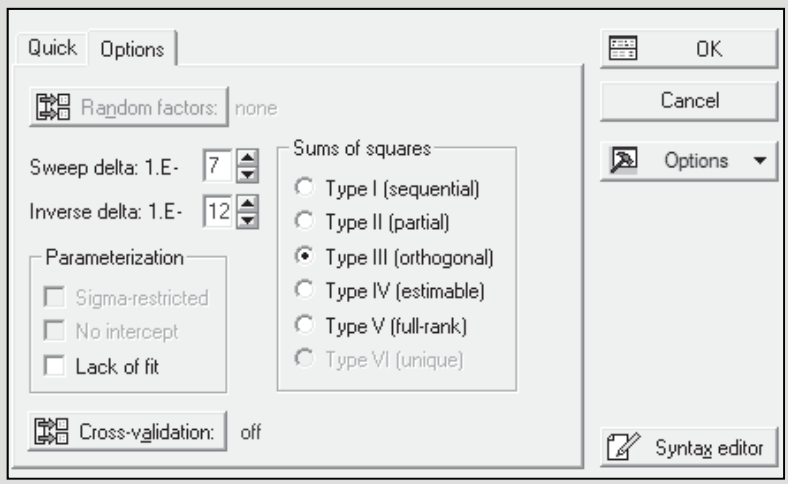
Paso 8. En la ventana del paso 6, se pulsa en «Between effects» y aparece esta nueva ventana donde se especifica que el factor «Turno» está anidado en el factor «Colegio», simplemente pulsando sobre «Colegio».



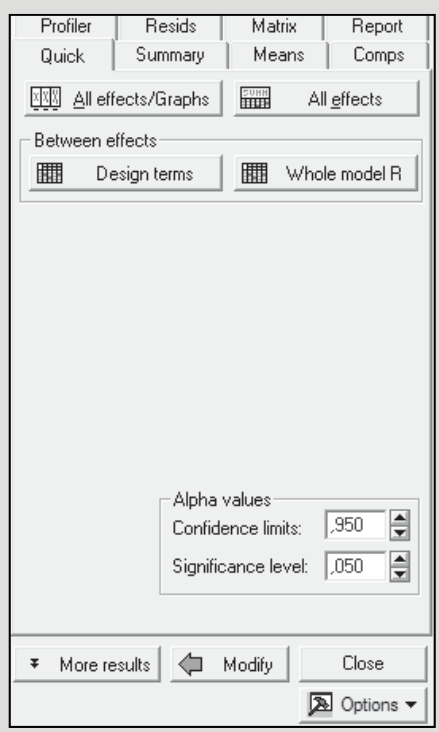
Paso 9. En la ventana del paso 6, en la pestaña «Options» aparece la siguiente ventana donde es posible elegir el modelo.

CUADRO IX.6. *(Continuación)*[STATISTICA]

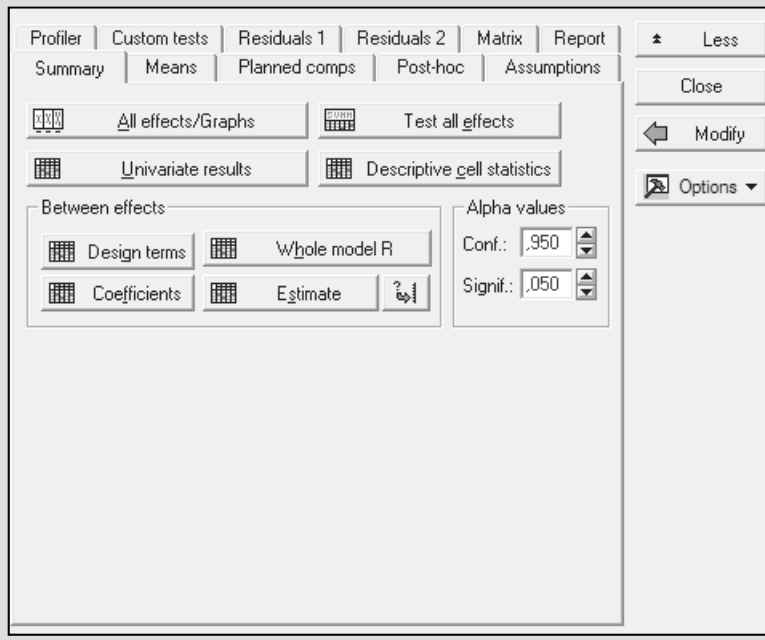
Se elige la opción «Type III (orthogonal)», adecuada en general para diferentes características de los datos y distintos diseños. También podría elegirse la opción «Type I», que funciona adecuadamente con modelos totalmente anidados.



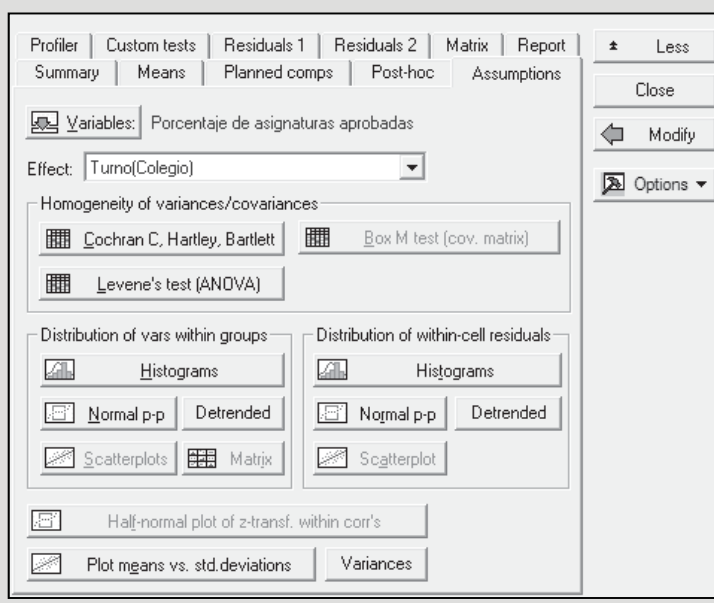
Paso 10. Al pulsar «OK» obtenemos la siguiente ventana donde pulsamos en «More results» para que aparezca la ventana extendida que se muestra en la siguiente página.



CUADRO IX.6. *(Continuación)* *STATISTICA*



Paso 11. En la pestaña «Assumptions» pulsamos «Levene´s test (ANOVA)».

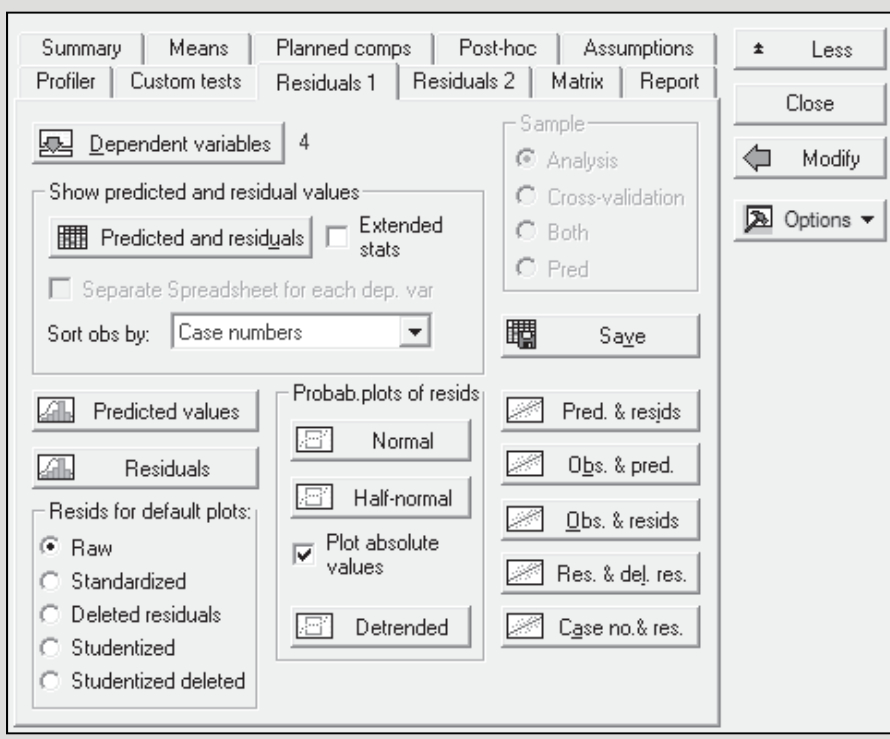


CUADRO IX.6. *(Continuación)*[STATISTICA]

Paso 12. Aparece la tabla que se muestra a continuación. Como la probabilidad es mayor que 0,05, aceptamos la hipótesis nula de homogeneidad de varianzas (p = 0,484). Por lo tanto, se cumple este requisito necesario para dar validez a los resultados que se obtienen con el ANOVA.

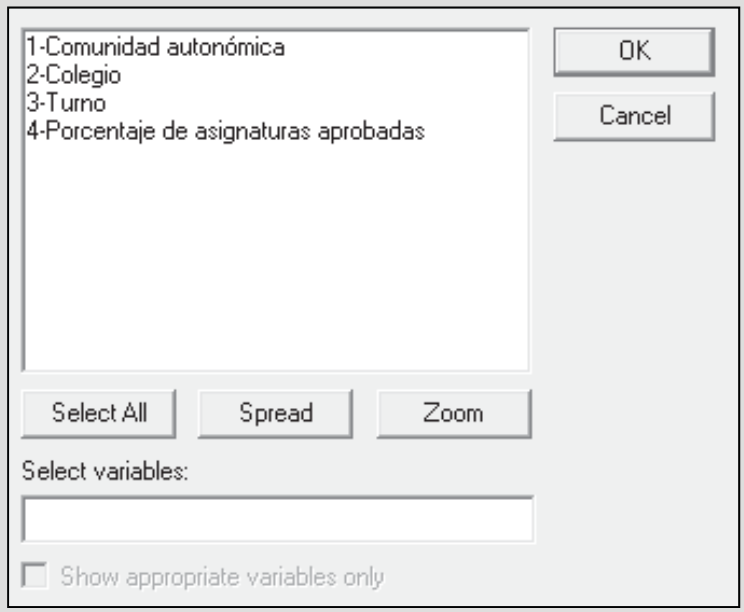
	Levene's Test for Homogeneity of Variances (Cuadro IX.6.sta) Effect: Turno(Colegio) Degrees of freedom for all F's: 5, 18 Include condition: v1="Madrid"			
	MS Effect	MS Error	F	p
Porcentaje de asignaturas aprobadas	2,946030	3,163571	0,931236	0,484010

Paso 13. Para comprobar la Normalidad, lo que se hace es ver si los residuos tienen distribución Normal. En la pestaña «Residuals 1» pulsamos «Save».



Paso 14. En la ventana que aparece, y se muestra en la siguiente página, pulsamos «OK» sin seleccionar ninguna variable y se obtiene la tabla que se muestra a continuación, en la cual la tercera columna son los residuos.

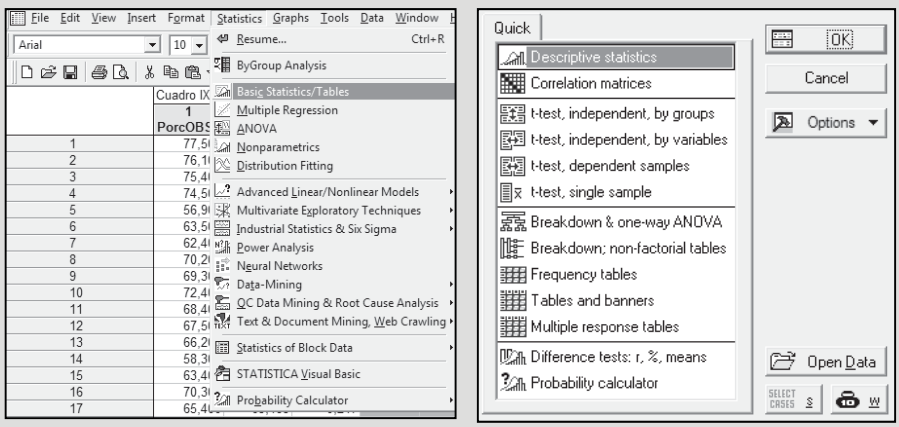
CUADRO IX.6. *(Continuación)*[STATISTICA]



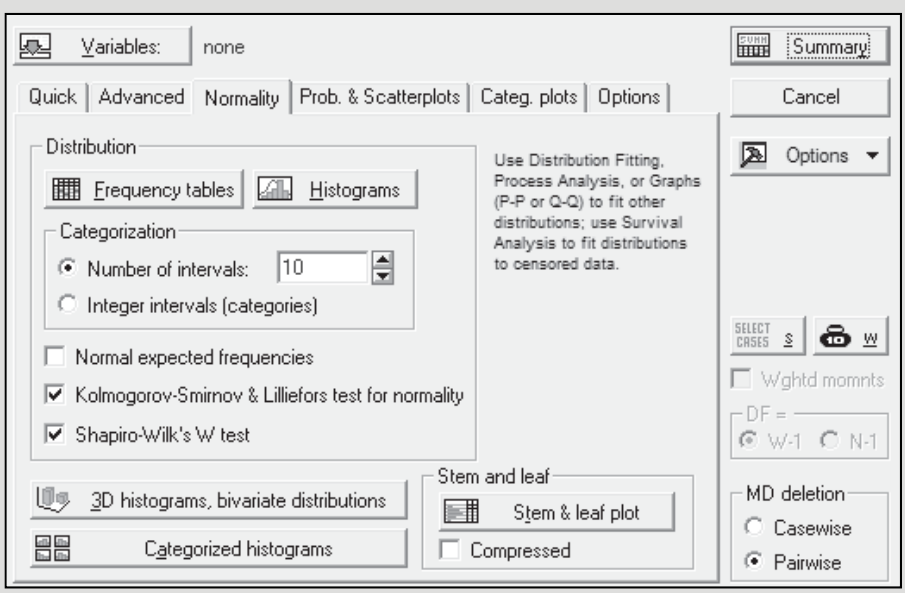
	Cuadro IX.6.sta		
	1 PorcOBS1	2 PorcPRE1	3 PorcRES1
1	77,500	75,875	1,625
2	76,100	75,875	0,225
3	75,400	75,875	-0,475
4	74,500	75,875	-1,375
5	56,900	60,933	-4,033
6	63,500	60,933	2,567
7	62,400	60,933	1,467
8	70,200	70,075	0,125
9	69,300	70,075	-0,775
10	72,400	70,075	2,325
11	68,400	70,075	-1,675
12	67,500	65,183	2,317
13	66,200	65,183	1,017
14	58,300	65,183	-6,883
15	63,400	65,183	-1,783
16	70,300	65,183	5,117
17	65,400	65,183	0,217

CUADRO IX.6. *(Continuación)*[STATISTICA]

Paso 15. Para determinar si los residuos tienen distribución Normal se selecciona «Statistics» ⇨ «Basic Statistics/Tables» ⇨ «Descriptive statistics».

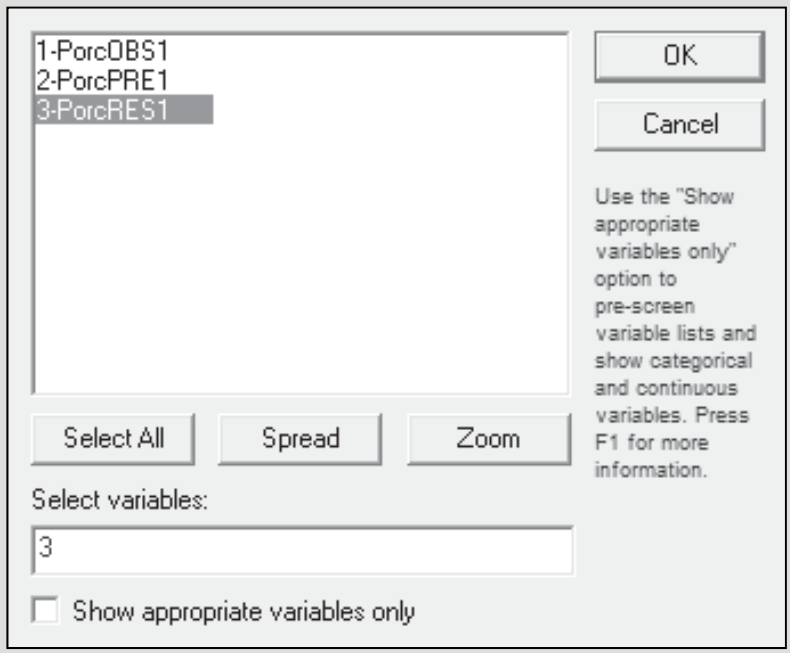


Paso 16. En la pestaña «Normality» marcamos los test «Kolmogorov-Smirnov & Lilliefors test for normality» y «Shapiro-Wilk´s test».

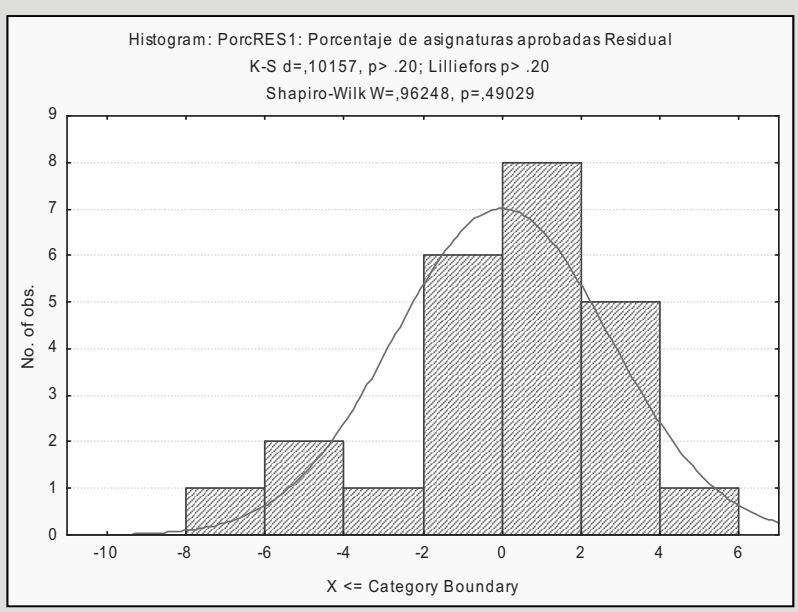


Paso 17. En la ventana anterior, en «Variables», aparece la ventana que se muestra en la siguiente página donde se selecciona «PorcRES1».

CUADRO IX.6. *(Continuación)*^{STATISTICA}

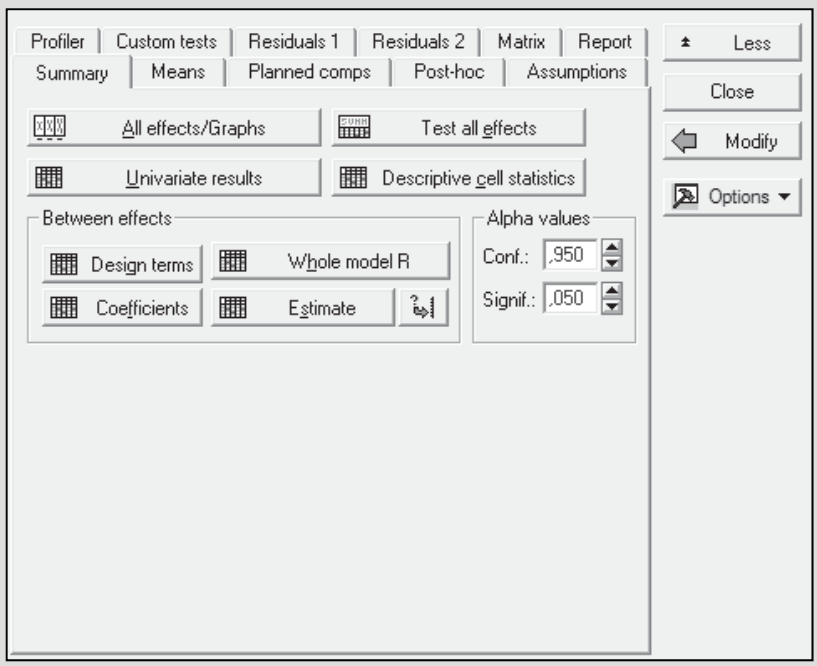


Paso 18. En la ventana del paso 16, al pulsar «Histograms» aparece el siguiente gráfico donde se observa que tanto el test Kolmogorov–Smirnov con la corrección de Lilliefors, como el de Shapiro-Wilk muestran valores de $p > 0{,}05$ y, por tanto, se acepta la hipótesis nula de bondad de ajuste y los residuos se ajustan a una distribución Normal.



CUADRO IX.6. *(Continuación)*^{STATISTICA}

Paso 19. En la ventana del menú principal del ANOVA, en la pestaña «Summary» pulsamos en «Univariate results» para obtener los resultados.



Paso 20. Los resultados muestran que no existen diferencias significativas entre los colegios (p = 0,882), pero si existen diferencias significativas entre el turno de mañana y el de la tarde en el porcentaje de asignaturas aprobadas por los alumnos (ANOVA, $F_{3,\ 18}$ = 17,4, p < 0,001).

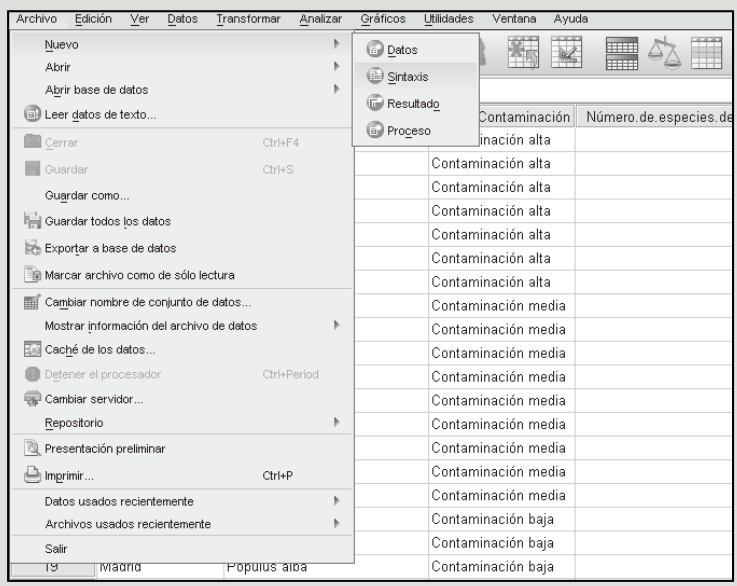
	Univariate Results for Each DV (Cuadro IX.6.sta) Over-parameterized model Type III decomposition Include condition: v1="Madrid"				
Effect	Degr. of Freedom	Porcentaje de asignaturas aprobadas SS	Porcentaje de asignaturas aprobadas MS	Porcentaje de asignaturas aprobadas F	Porcentaje de asignaturas aprobadas p
Intercept	1	105139,5	105139,5	11016,79	0,000000
Colegio	2	2,4	1,2	0,13	0,882421
Turno(Colegio)	3	501,7	167,2	17,52	0,000014
Error	18	171,8	9,5		
Total	23	697,4			

CUADRO IX.6. *(Continuación)*[SPSS]

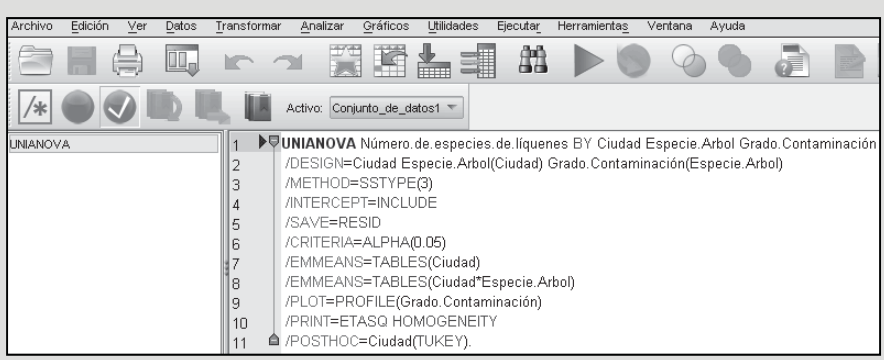
Aplicación con SPSS.

EJEMPLO. Vamos a utilizar el mismo ejemplo que con R. El objetivo es determinar si el número de especies de líquenes presentes en dos especies de árboles, *Populus alba* (Chopo blanco) y *Aesculus hippocastanum* (Castaño de indias), en tres ciudades de España (Madrid, Barcelona y Sevilla) varía en función del grado de contaminación (poca, media y alta contaminación). Los datos están en el archivo **Cuadro IX.6.sav**.

Paso 1. Para realizar un ANOVA anidado es mejor usar el editor de sintaxis que el sistema de ventanas de SPSS. Una vez abierto el archivo **Cuadro IX.6.sav** vamos a «Archivo» ⇨ «Abrir» ⇨ «Sintaxis…» y abrimos el archivo **Cuadro IX.6.sps**.



Paso 2. El archivo **Cuadro IX.6.sps** tiene las instrucciones para realizar el ANOVA anidado, las cuales se explican a continuación.



CUADRO IX.6. *(Continuación)*[SPSS]

Paso 3. En primer lugar hay que definir el tipo de análisis que se va a realizar, en este caso «UNIANOVA», la variable dependiente «Número.de.especies.de.líquenes» y los factores «Ciudad», «Especie.Arbol» y «Grado.Contaminación», separando la variable dependiente de los factores por «BY».

UNIANOVA Número.de.especies.de.líquenes BY Ciudad Especie.Arbol Grado.Contaminación

Paso 4. Como el diseño es de un ANOVA anidado, hay que especificar cuales son los factores anidados. En este caso «Ciudad» es el factor principal que no está anidado en ningún otro factor, la especie de árbol estaría anidada dentro de «Ciudad» y, por ello, se especificaría como «Especie.Arbol(Ciudad)» y, por último, el grado de contaminación es un factor que estaría anidado dentro de la especie de árbol «Grado.Contaminación(Especie.Arbol(Ciudad))».

/DESIGN=Ciudad Especie.Arbol(Ciudad)
Grado.Contaminación(Especie.Arbol)

Paso 5. Se elige la opción sumas de cuadrados tipo III, adecuada en general para diferentes características de los datos y distintos diseños. También podría elegirse el tipo I, que funciona adecuadamente con modelos totalmente anidados.

/METHOD=SSTYPE(3)

Paso 6. Como en cualquier análisis de varianza es posible incluir o no la constante en el modelo, aunque lo normal es incluirla.

/INTERCEPT=INCLUDE

Paso 7. Es importante guardar los residuos para más tarde analizar si tienen una distribución Normal.

/SAVE=RESID

Paso 8. Normalmente se utiliza 0,05 como nivel de significación.

/CRITERIA=ALPHA(0.05)

Paso 9. La siguiente instrucción permite obtener datos descriptivos de los factores, como la media y el error. Se especifica para el factor «Ciudad» y para la interacción de «Ciudad» y «Especie.Arbol».

/EMMEANS=TABLES(Ciudad)
/EMMEANS=TABLES(Ciudad*Especie.Arbol)

Paso 10. La siguiente instrucción permite representar los valores medios del número de especies de líquenes para cada nivel de contaminación.

/PLOT=PROFILE(Grado.Contaminación)

CUADRO IX.6. *(Continuación)*[SPSS]

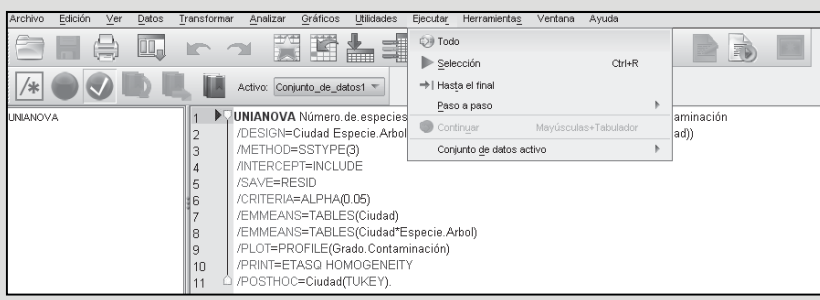
Paso 11. La instrucción «ETASQ» permite obtener el estadístico Eta al cuadrado parcial, el cual cuantifica la importancia de cada factor sobre la variable dependiente. La instrucción «HOMOGENEITY» permite obtener los resultados del test de Levene para determinar si existe homogeneidad de varianzas, el otro requisito del análisis.

```
/PRINT=ETASQ HOMOGENEITY
```

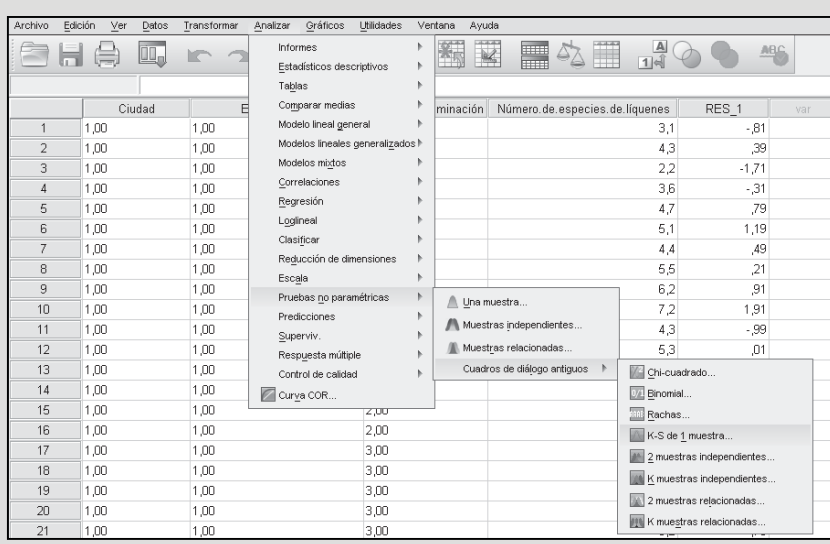
Paso 12. Por último, como en el factor «Ciudad» hay más de dos niveles, en concreto tres ciudades, es posible hacer pruebas *post hoc* para ver las diferencias entre cada par de ciudades. Existen pruebas para cuando las varianzas son iguales o son diferentes. Para varianzas iguales, una prueba que se utiliza frecuentemente es el test «Tukey».

```
/POSTHOC=Ciudad(TUKEY).
```

Paso 13. Para realizar el análisis se selecciona «Ejecutar» ⇨ «Todo».

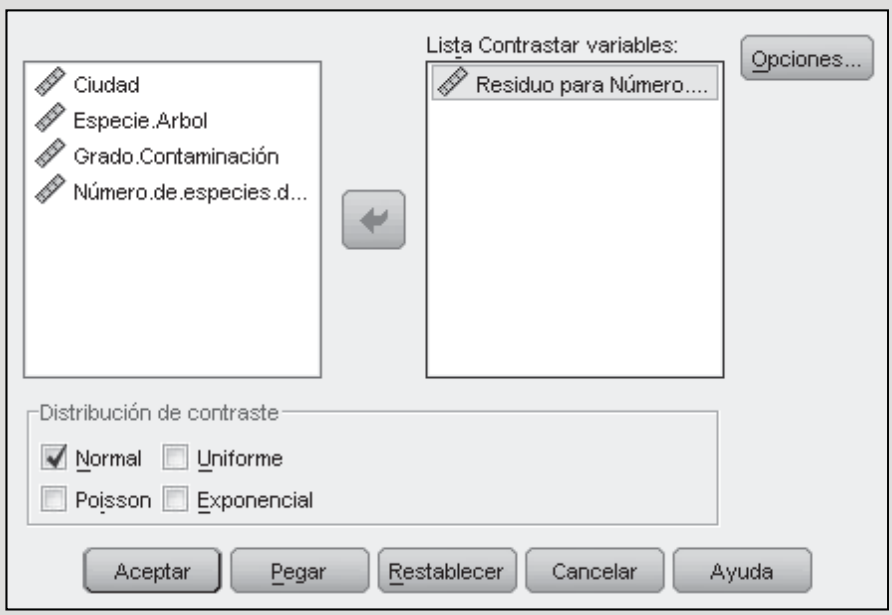


Paso 14. En primer lugar hay que comprobar el requisito de que los residuos tengan una distribución Normal. Para ello se selecciona «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «K-S de 1 muestra…».



CUADRO IX.6. *(Continuación)*[SPSS]

Paso 15. En «Lista Contrastar variables:» se introduce el residuo y en «Distribución de contraste» se marca «Normal».



Paso 16. El test de Kolmogorov-Smirnov muestra que los residuos tienen una distribución Normal (p = 0,817). Si los residuos no tuviesen una distribución Normal, sería necesario transformar los datos como se explicó en los pasos 12 a 17 del Cuadro IX.3.

		Residuo para Número.de.especies.de.líquenes
N		166
Parámetros normales[a,b]	Media	,0000
	Desviación típica	1,62897
Diferencias más extremas	Absoluta	,049
	Positiva	,049
	Negativa	-,033
Z de Kolmogorov-Smirnov		,634
Sig. asintót. (bilateral)		,817

Paso 17. Los resultados del test de Levene, que se ven en la siguiente tabla, muestran que también existe homogeneidad de varianzas (p = 0,102) y, por tanto, se cumple el otro requisito del ANOVA.

CUADRO IX.6. *(Continuación)*[SPSS]

Contraste de Levene sobre la igualdad de las varianzas error[a]

Variable dependiente:Número.de.especies.de.líquenes

F	gl1	gl2	Sig.
1,501	17	148	,102

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intersección + Ciudad + Especie.Arbol(Ciudad) + Grado.Contaminación(Especie.Arbol)

Paso 18. Los resultados del ANOVA se muestran en la siguiente tabla. Existen diferencias significativas entre las ciudades en el número de especies de líquenes (ANOVA, $F_{2,\ 156} = 12{,}651$, $p < 0{,}001$), pero no existen diferencias entre las dos especies de árboles (ANOVA, $F_{2,\ 156} = 0{,}249$, $p = 0{,}78$). En lo que se refiere al objetivo del ejemplo, existen diferencias significativas en el número de especies de líquenes observadas en función del grado de contaminación (ANOVA, $F_{4,\ 156} = 94{,}347$, $p < 0{,}001$).

Pruebas de los efectos inter-sujetos

Variable dependiente:Número.de.especies.de.líquenes

Origen	Suma de cuadrados tipo III	gl	Media cuadrática	F	Sig.	Eta al cuadrado parcial
Modelo corregido	1159,880[a]	9	128,876	45,918	,000	,726
Intersección	8046,051	1	8046,051	2866,807	,000	,948
Ciudad	71,012	2	35,506	12,651	,000	,140
Especie.Arbol(Ciudad)	1,395	2	,698	,249	,780	,003
Grado.Contaminación(Especie.Arbol)	1059,188	4	264,797	94,347	,000	,708
Error	437,833	156	2,807			
Total	10625,820	166				
Total corregida	1597,714	165				

a. R cuadrado = ,726 (R cuadrado corregida = ,710)

El estadístico Eta cuadrado parcial informa sobre el porcentaje de la variabilidad residual de la variable dependiente (no explicada ya por otras variables) que es explicada por cada factor, en un rango de 0 a 1. Se observa que el Eta cuadrado parcial en el factor anidado «Grado.Contaminación(Especie.Arbol)» es alto (0,708). Por tanto, el grado de contaminación explica un alto porcentaje de la variación observada en el número de líquenes en los árboles.

CUADRO IX.6. *(Continuación)*[SPSS]

Paso 19. Las pruebas *post hoc* del test de Tukey para el factor «Ciudad» muestran que hay diferencias significativas entre Madrid y las otras dos ciudades ($p < 0{,}001$), pero no hay diferencias entre Barcelona y Sevilla ($p = 0{,}997$).

Comparaciones múltiples

Número.de.especies.de.líquenes

DHS de Tukey

(I)Ciudad	(J)Ciudad	Diferencia de medias (I-J)	Error típ.	Sig.	Intervalo de confianza 95%	
					Límite inferior	Límite superior
Madrid	Barcelona	-1,437*	,3168	,000	-2,187	-,687
	Sevilla	-1,413*	,3224	,000	-2,176	-,650
Barcelona	Madrid	1,437*	,3168	,000	,687	2,187
	Sevilla	,024	,3168	,997	-,726	,774
Sevilla	Madrid	1,413*	,3224	,000	,650	2,176
	Barcelona	-,024	,3168	,997	-,774	,726

Basadas en las medias observadas.

El término de error es la media cuadrática(Error) = 2,807.

*. La diferencia de medias es significativa al nivel 0,05.

Por lo tanto, si bien el ANOVA mostró diferencias entre ciudades (véase paso 18), ello se debe a que los valores en Madrid son muy diferentes en comparación con las otras dos ciudades. Estas diferencias entre Madrid y las otras ciudades se pueden ver en la siguiente tabla, que muestra los valores medios y el error obtenido en cada ciudad. Se observa como en Madrid el número medio de especies de líquenes por árbol es 6,1, mientras que es 7,55 en Barcelona y 7,53 en Sevilla.

1. Ciudad

Variable dependiente:Número.de.especies.de.líquenes

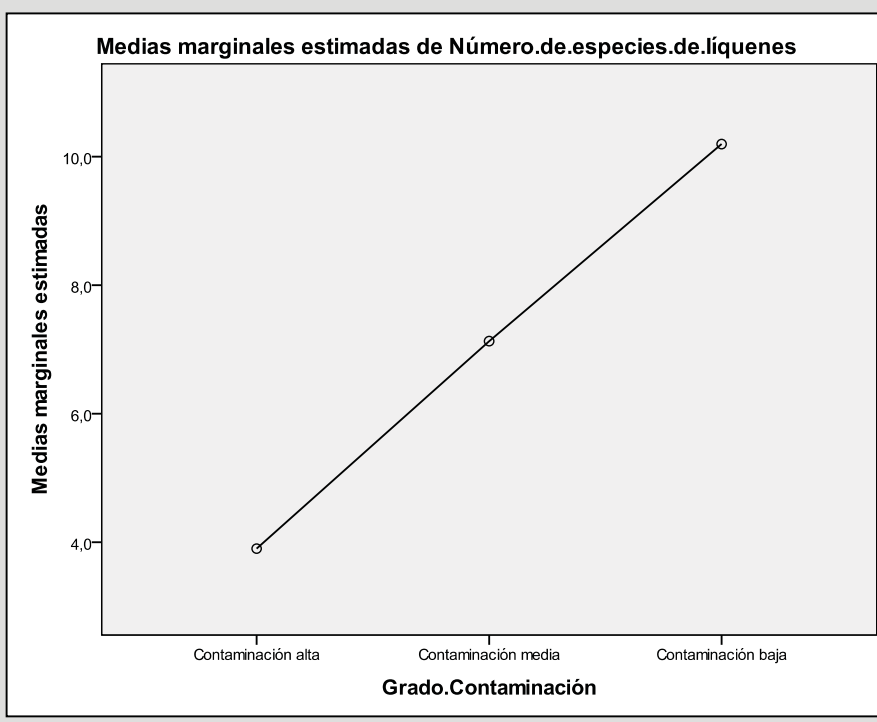
Ciudad	Media	Error típ.	Intervalo de confianza 95%	
			Límite inferior	Límite superior
Madrid	6,141	,229	5,688	6,593
Barcelona	7,551	,222	7,112	7,989
Sevilla	7,531	,231	7,075	7,986

CUADRO IX.6. *(Continuación)*[SPSS]

Paso 20. El hecho de que el ANOVA no mostrara diferencias significativas entre las dos especies de árbol también se aprecia claramente en la siguiente tabla, donde se observa que, en cada ciudad, el número medio de especies de líquenes en *Populus alba* y en *Aesculus hippocastanum* es muy parecido.

Ciudad	Especie.Arbol	Media	Error típ.	Intervalo de confianza 95%	
				Límite inferior	Límite superior
Madrid	Populus alba	5,972	,330	5,321	6,624
	Aesculus hippocastanum	6,309	,317	5,682	6,936
Barcelona	Populus alba	7,168	,302	6,572	7,764
	Aesculus hippocastanum	7,934	,326	7,291	8,577
Sevilla	Populus alba	7,309	,335	6,648	7,970
	Aesculus hippocastanum	7,752	,317	7,125	8,379

Paso 21. Por último, el siguiente gráfico muestra claramente que el número de especies de líquenes es mayor a medida que la contaminación es más baja.



IX.2. Pruebas no paramétricas

En muchas ocasiones no podemos asumir los supuestos sobre normalidad e igualdad de varianzas en nuestros datos, por tanto es necesario utilizar la llamada estadística no paramétrica o métodos de distribución libre. Igualmente, podemos utilizar los tests no paramétricos cuando los datos de la población están en forma de rango, cuando el procedimiento de obtención de la muestra impide asimilarla con una sucesión de valores cuantitativos (Pérez 2004).

Tienen como ventaja que son de aplicación más general que las paramétricas (a las variables distribuidas normalmente también se les puede aplicar estos contrastes), ya que no exigen ninguna condición sobre el tipo de distribución. Sin embargo, son menos sensibles a la detección de diferencias que las pruebas paramétricas, aunque de forma general se puede decir que la coincidencia entre los resultados obtenidos entre las pruebas paramétricas y las no paramétricas es superior al 90%.

Dentro de los contrastes de homogeneidad no paramétricos podemos distinguir dos tipos diferentes en función de la naturaleza de las muestras. Tendríamos muestras independientes cuando los elementos sean seleccionados aleatoriamente de forma separada o no relacionada (por ejemplo, medida del nivel de azúcar en hombres y mujeres), y muestras dependientes, cuando las mediciones si están relacionadas, generalmente porque corresponden al mismo individuo (por ejemplo, medida del nivel de azúcar en diabéticos, antes y después de un tratamiento).

IX.2.1. CONTRASTES PARA DOS MUESTRAS INDEPENDIENTES

Son los contrastes análogos al contraste paramétrico *t* de Student. Existen tres pruebas principales que tratan de determinar si las diferencias observadas entre dos muestras son debidas al azar o si pertenecen a dos poblaciones distintas. En todas ellas, los datos se ordenan de forma ascendente sin considerar su grupo de pertenencia y, en función de su posición, se les asigna un número de orden o rango.

IX.2.1.1. Prueba *U* de Mann-Whitney

Es una prueba que compara la tendencia central de dos muestras, que no es necesario que tengan el mismo tamaño, partiendo de la hipótesis nula de que en ambas muestras la medida central es la misma.

Es un contraste muy utilizado, aunque en el caso de que se quiera medir no sólo diferencias en la tendencia central de los datos sino en otras características como, por ejemplo, asimetría o dispersión de los datos, es más adecuado utilizar el test de Wald-Wolfowitz.

Una premisa de esta prueba es que los valores de cada una de las muestras son diferentes, es decir, que no hay solapamiento de datos (valores repetidos). En caso de que existan solapamientos, el contraste se vuelve más conservador (es más difícil encontrar diferencias entre muestras). Si el número de solapamientos es muy grande, sobre todo en muestras pequeñas y con variables continuas, se aconseja la utilización del test de Kolmogorov-Smirnov para dos muestras, aunque existe una corrección para solapamientos del propio estadígrafo.

A las muestras ordenadas de menor a mayor se les asigna un rango de forma ascendente (1,2,3,…..), habiendo tantos rangos como datos. En caso de solapamiento o empate se suman los rangos implicados en el mismo y se divide por el número de datos solapados. Un ejemplo de asignación de rangos aparece en la Tabla IX.2.

Tabla IX.2. Proceso para la asignación de rangos a un grupo de datos.

Valores muestra 1	1	3	6	5	8					
Valores muestra 2	2	4	9	8	1					
Valores ordenados	1	1	2	3	4	5	6	8	8	9
Número de rango	1,5	1,5	3	4	5	6	7	8,5	8,5	10

El desarrollo completo de esta prueba aparece en Sokal & Rohlf (1981). Consiste en calcular el estadístico *U* con el que es posible conocer si la media de los rangos es significativamente diferente entre las dos muestras. Si el número total de datos es inferior a 20, se utiliza una tabla para la determinación de valores críticos de *U*. En cambio, si el número total de datos es superior a 20, se puede utilizar una aproximación a la Normal a partir del estadístico *U*, que nos indica el valor *Z* de dicha Normal y su probabilidad.

En el Cuadro IX.7 se explica la obtención de este estadístico de contraste.

IX.2.1.2. Test de rachas de Wald-Wolfowitz

Es una prueba que compara la distribución de dos muestras, que no es necesario que tengan el mismo tamaño, partiendo de la hipótesis nula de que la distribución de rangos es aleatoria, es decir, las muestras son homogéneas.

Los datos de las dos muestras se ordenan conjuntamente de forma ascendente y a cada grupo de valores se le asigna un valor de racha. Las rachas son secuencias de valores del mismo grupo, cuando las muestras han sido ordenadas. Un ejemplo de ordenación en rachas aparece en la Tabla IX.3.

Tabla IX.3. Proceso para la asignación de rachas a un grupo de datos

Valores muestra 1	1	2	6	5	8					
Valores muestra 2	2	4	6	8	2					
Valores ordenados	1	2	2	2	4	5	6	6	8	8
Número de racha	1	1	2	2	2	3	3	4	4	5

Una descripción completa del estadístico aparece en el apartado 25.6 del libro *Biostatistical Analysis* (Zar 1999).

Si dos muestras presentan una misma distribución cabe esperar que, en el ordenamiento de datos de menor a mayor, ambas muestras estén muy mezcladas (aleatorizadas), es decir, que el número de rachas sea alto. Con una tabla de contraste se determina si el número de rachas encontradas es significativamente grande y se puede suponer la aleatoriedad de los rangos y, por tanto, que no hay diferencias entre las muestras. Si el número de datos es grande, al igual que sucedía con el contraste de Mann-Whitney, este estadístico se ajusta a una Normal y se puede analizar la homogeneidad de las muestras calculando el estadístico *Z*.

Este estadístico es menos potente que el contraste de Mann-Whitney, pero tiene la ventaja de que detecta diferencias no sólo de tendencia central sino de dispersión y asimetría.

En el Cuadro IX.7 se describen los pasos para la obtención de este estadístico.

IX.2.1.3. Prueba de Kolmogorov-Smirnov para dos muestras

Tiene como ventaja que mide las diferencias existentes entre las frecuencias relativas acumuladas de las dos muestras, por lo que se detectan diferencias no sólo en la tendencia central, sino también en la dispersión y forma (simetría, apuntamiento) de las muestras.

Cuando el tamaño de la muestra es grande, es mejor el estadístico *U* de Mann-Whitney.

Esta prueba estadística se usa para contrastar la H_0 de que las dos muestras proceden de una misma población, por tanto exige comparar dos funciones de distribución muestral, observando la diferencia máxima entre ellas. El desarrollo completo del cálculo de este estadístico se encuentra en Sokal & Rohlf (1981), en el que los pasos a seguir son:

1. Ordenar las observaciones de cada muestra y distribuirlas en clases.
2. Determinar las frecuencias acumuladas de las clases en cada muestra.
3. Buscar la máxima diferencia entre la frecuencia acumulada de ambas muestras.
4. Determinar los valores críticos de D_{max} en una tabla.
5. Si D_{max} es mayor que el valor crítico, se rechaza la hipótesis nula de que las muestras proceden de la misma población.

Un ejemplo de la utilización de este test aparece en el Cuadro IX.7.

CUADRO IX.7. Contrastes de homogeneidad no paramétricos para dos muestras independientes

Aplicación con R.

EJEMPLO. En el archivo **Cuadro IX.7.V.csv** hay datos de la concentración de nitrito, nitrato y amonio (μM l^{-1}) en lagos de la amazonia colombiana en dos meses diferentes. El objetivo es determinar si existen diferencias en la concentración de nutrientes entre cada dos lagos, para cada uno de los meses.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro IX.7.R**, que se muestra en la siguiente ventana. Sólo es necesario especificar lo que se solicita en la sección modificable por el usuario. Para que este *script* funcione correctamente es necesario tener instalados los paquetes "adehabitat", "gpclib" y "ade4", o realizar la instalación como se explica en el Apéndice II.

CUADRO IX.7. *(Continuación)*[R]

```
#Cuadro IX.7
#Contrastes de homogeneidad no paramétricos para dos muestras independientes.

################################################
#Sección modificable por el usuario
################################################

#Lectura de datos.
datos<-read.csv2("Cuadro IX.7.V.csv",header=TRUE,encoding="latin1")

#Selección variable dependiente.
varInteres<-c("Nitrito")
#varInteres<-c("Nitrato")
#varInteres<-c("Amonio")

#Variable que corresponde al factor.
varFactor<-c("Lago")
#varFactor<-c("Mes")

#Selección de grupos a comparar en el factor.
varComparar1<-c("Correo")
varComparar2<-c("Tarapoto")
#varComparar1<-NULL

##Variables de selección.
#Los elementos de las listas varSeleccion pueden modificarse en función de nuestros intereses.
#Pueden escogerse hasta dos grupos en cada variable de selección, en un número igual o inferior al nº total de grupos - 1
varSeleccion1<-list(c("Mes","Febrero"))
#varSeleccion1<-NULL
#varSeleccion2<-list(c("Lago","Correo","Tarapoto"))
varSeleccion2<-NULL

#Archivo de salida.
ArchivodeSalida<-"Salida Cuadro IX.7.txt"
```

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro IX.7.V.csv**.

datos<-read.csv2("Cuadro IX.7.V.csv",header=TRUE,encoding="latin1")

Paso 2. Se selecciona la variable dependiente, en este caso se realiza el ejemplo para la concentración de nitritos.

varInteres<-c("Nitrito")
#varInteres<-c("Nitrato")
#varInteres<-c("Amonio")

Paso 3. Se indica el factor con el que se va trabajar, en el ejemplo se comparan los lagos.

varFactor<-c("Lago")
#varFactor<-c("Mes")

Paso 4. A continuación se indica cuales son los grupos a comparar, los lagos Tarapoto y Correo. Si solo existieran 2 niveles dentro del factor, no sería necesario especificar los grupos a comparar y simplemente se activaría la acción «NULL».

varComparar1<-c("Correo")
varComparar2<-c("Tarapoto")
#varComparar1<-NULL

CUADRO IX.7. *(Continuación)*[R]

Paso 5. Es posible seleccionar grupos en dos variables diferentes (un máximo de dos grupos en cada variable con este *script*). En el ejemplo, como se quiere comparar sólo el mes de febrero, se especifica el nombre de la variable selección y los grupos que se seleccionan dentro de ella. Por último, se pone el símbolo # delante del «NULL». Se observa que sólo se utiliza la variable selección 1, ya que en la 2 está activada la instrucción «NULL».

```
varSeleccion1<-list(c("Mes","Febrero"))
#varSeleccion1<-NULL
#varSeleccion2<-list(c("Lago","Correo","Tarapoto"))
varSeleccion2<-NULL
```

Paso 6. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
ArchivodeSalida<-"Salida Cuadro IX.7.txt"
```

Paso 7. El test de Wilcoxon de suma de rangos (que como siempre da una probabilidad idéntica a la prueba *U* de Mann-Whitney, es el que utiliza R) con una $p = 0{,}03$ rechaza la hipótesis de igualdad, mientras que el de Kolmogorov-Smirnov con una $p = 0{,}142$ acepta la hipótesis. En este tipo de situaciones es conveniente seguir el resultado más conservador y, por tanto, se concluye que existen diferencias significativas en la concentración de nitritos, en el mes de febrero, entre los lagos Tarapoto y Correo ($p = 0{,}03$). Es importante mencionar que cuando existen empates, es decir, existen valores iguales en los grupos a comparar, R no calcula el valor exacto de probabilidad, ni en el test de Wilcoxon de suma de rangos ni en el test de Kolmogorov-Smirnov y, por ello, se pueden observar pequeñas diferencias con el resultado de SPSS y STATISTICA.

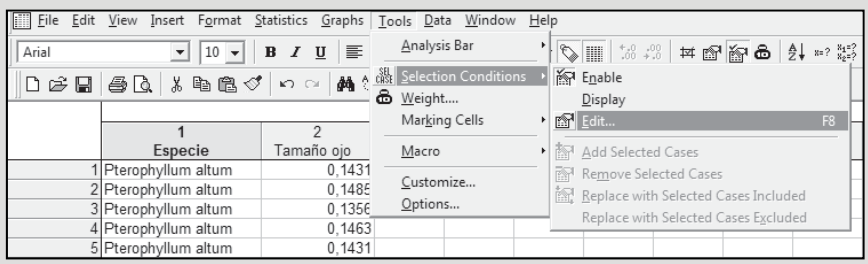
```
[[1]]

        Wilcoxon rank sum test with continuity correction

data:  datos3a$valores and datos3b$valores
W = 17.5, p-value = 0.03017
alternative hypothesis: true location shift is not equal to 0


[[2]]

        Two-sample Kolmogorov-Smirnov test

data:  datos3a$valores and datos3b$valores
D = 0.5341, p-value = 0.1423
alternative hypothesis: two-sided
```

CUADRO IX.7. *(Continuación)*[STATISTICA]

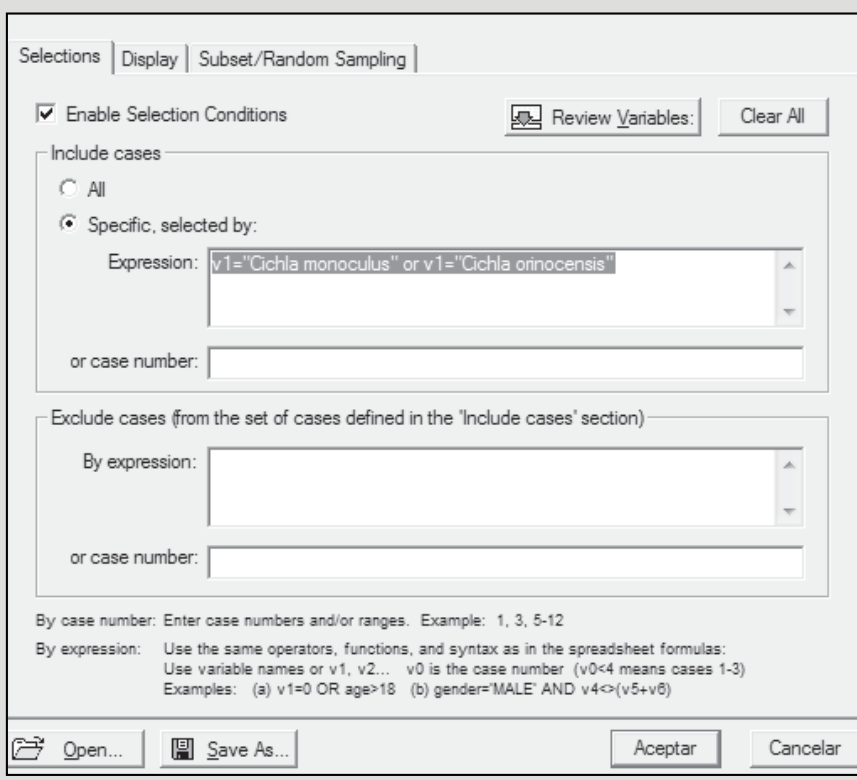
Aplicación con STATISTICA.

EJEMPLO. En el archivo **Cuadro IX.7.sta** hay datos de la relación tamaño del ojo/longitud del cuerpo de varias especies de perciformes de la familia Cichlidae. El objetivo es determinar si existen diferencias entre dos especies del mismo género: *Cichla monoculus* y *Cichla orinocensis*.

Paso 1. En primer lugar se seleccionan las especies que se quieren estudiar. Para ello vamos a «Tools» ⇨ «Selection Conditions» ⇨ «Edit».

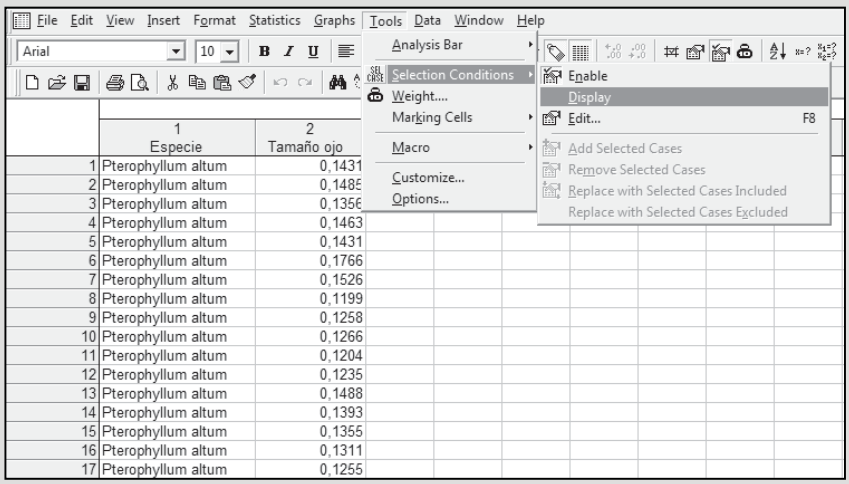


Paso 2. En «Specific, selected by:» se pone la expresión «v1="Cichla monoculus" or v1="Cichla orinocensis"».


Selections
Display
Subset/Random Sampling
Enable Selection Conditions
Review Variables:
Clear All
Include cases
All
Specific, selected by:
Expression: v1="Cichla monoculus" or v1="Cichla orinocensis"
or case number:
Exclude cases (from the set of cases defined in the 'Include cases' section)
By expression:
or case number:
By case number: Enter case numbers and/or ranges. Example: 1, 3, 5-12
By expression: Use the same operators, functions, and syntax as in the spreadsheet formulas:
Use variable names or v1, v2... v0 is the case number (v0<4 means cases 1-3)
Examples: (a) v1=0 OR age>18 (b) gender='MALE' AND v4<>(v5+v6)
Open...
Save As...
Aceptar
Cancelar

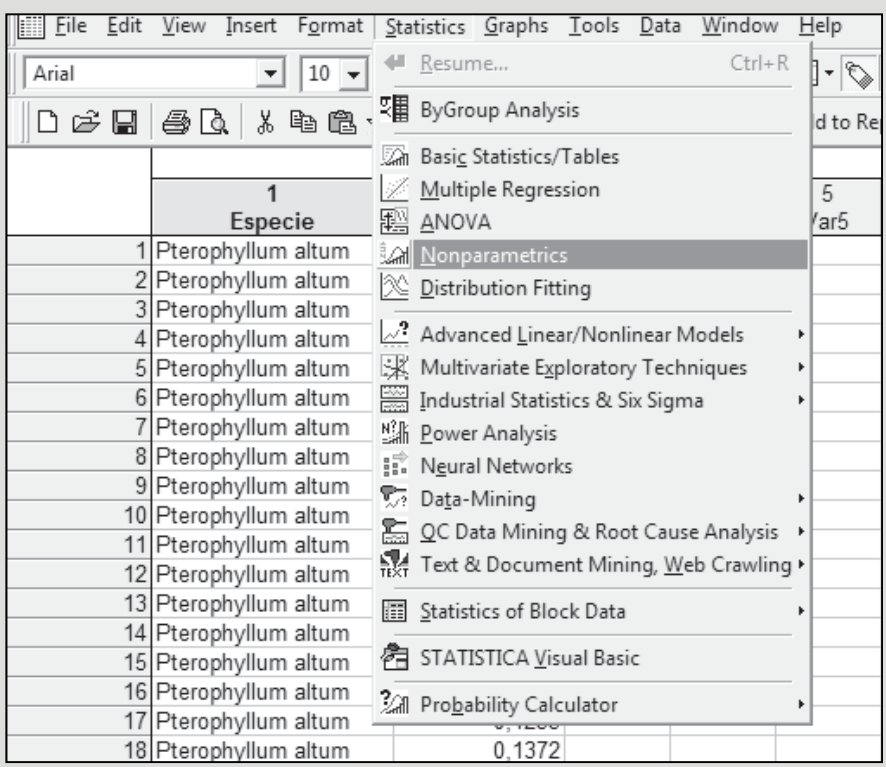
CUADRO IX.7. *(Continuación)*[STATISTICA]

Paso 3. Para asegurarnos de que se han seleccionado los grupos correctos se va a «Tools» ⇨ «Selection Conditions» ⇨ «Display». Las filas seleccionadas aparecerán en cursiva.



	1 Especie	2 Tamaño ojo
1	Pterophyllum altum	0,1431
2	Pterophyllum altum	0,1485
3	Pterophyllum altum	0,1356
4	Pterophyllum altum	0,1463
5	Pterophyllum altum	0,1431
6	Pterophyllum altum	0,1766
7	Pterophyllum altum	0,1526
8	Pterophyllum altum	0,1199
9	Pterophyllum altum	0,1258
10	Pterophyllum altum	0,1266
11	Pterophyllum altum	0,1204
12	Pterophyllum altum	0,1235
13	Pterophyllum altum	0,1488
14	Pterophyllum altum	0,1393
15	Pterophyllum altum	0,1355
16	Pterophyllum altum	0,1311
17	Pterophyllum altum	0,1255

Paso 4. Se selecciona «Statistics» ⇨ «Nonparametrics».

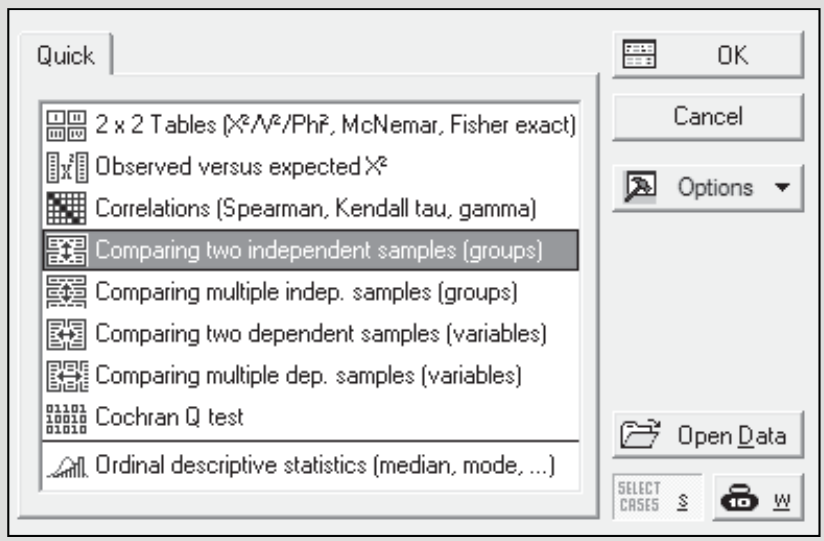


	1 Especie
1	Pterophyllum altum
2	Pterophyllum altum
3	Pterophyllum altum
4	Pterophyllum altum
5	Pterophyllum altum
6	Pterophyllum altum
7	Pterophyllum altum
8	Pterophyllum altum
9	Pterophyllum altum
10	Pterophyllum altum
11	Pterophyllum altum
12	Pterophyllum altum
13	Pterophyllum altum
14	Pterophyllum altum
15	Pterophyllum altum
16	Pterophyllum altum
17	Pterophyllum altum
18	Pterophyllum altum

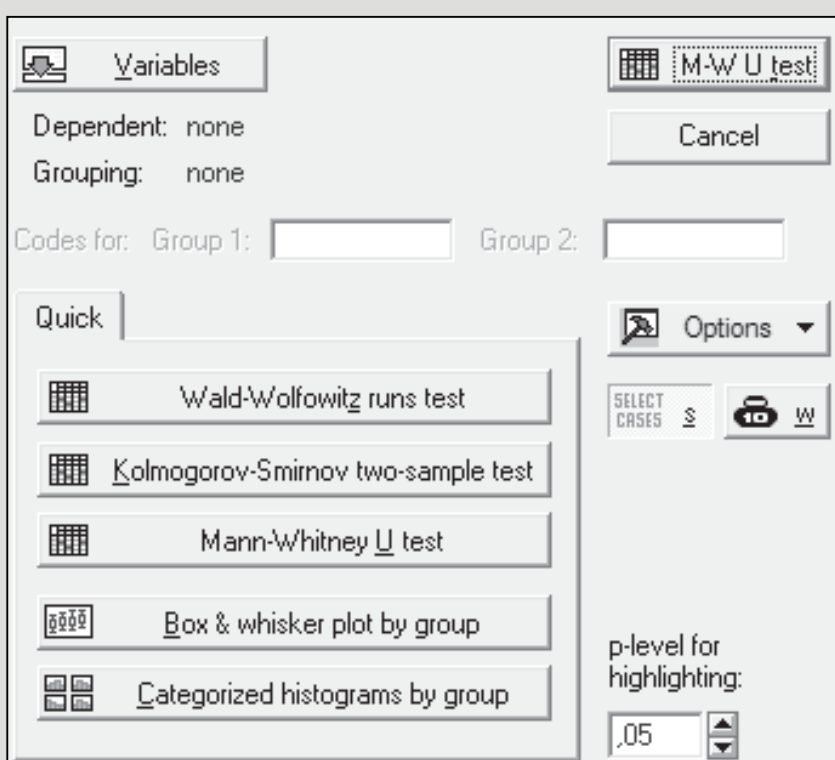
0,1372

CUADRO IX.7. *(Continuación)*[STATISTICA]

Paso 5. A continuación seleccionamos «Comparing two independent samples (groups)».

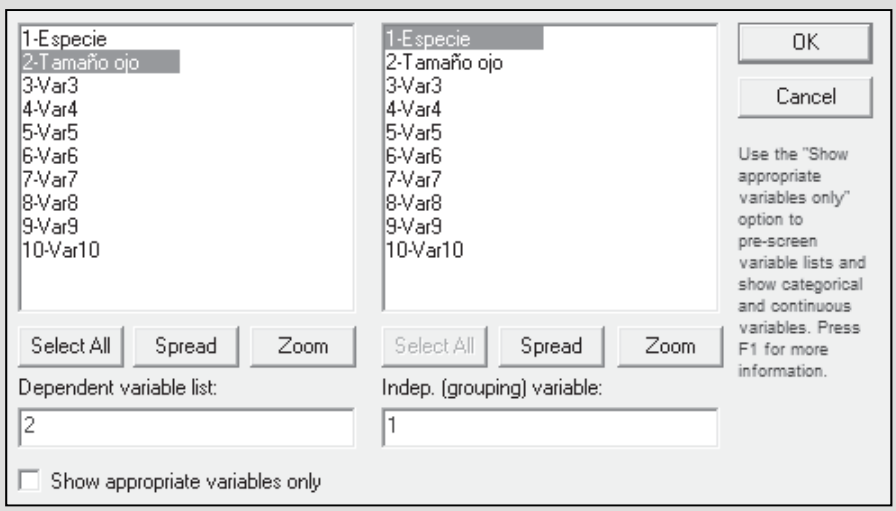


Paso 6. Pulsamos en «Variables».

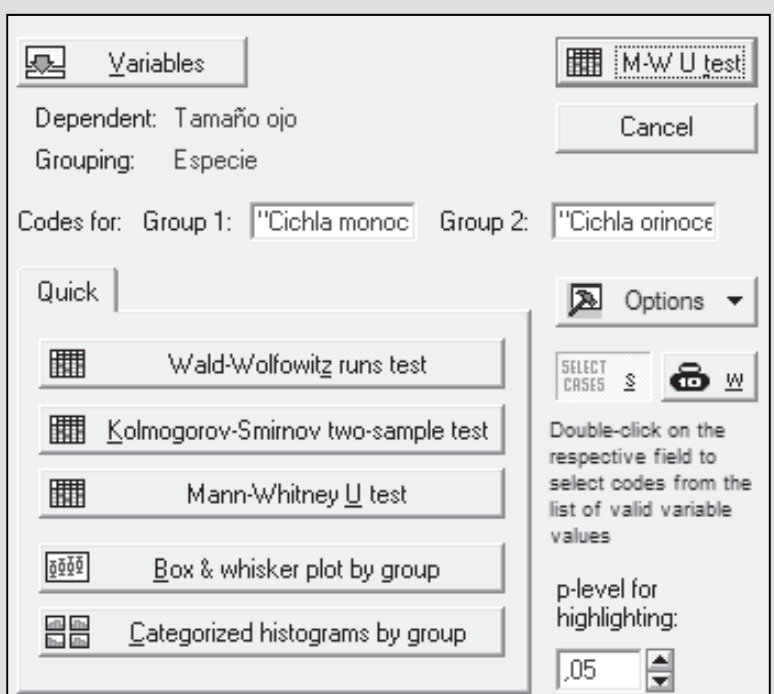


CUADRO IX.7. *(Continuación)*[STATISTICA]

Paso 7. Aparece la siguiente ventana, donde a la izquierda se indica la variable que contiene las mediciones «Tamaño ojo» y a la derecha la variable que indica el grupo de pertenencia «Especie».



Paso 8. En la ventana siguiente pulsamos sobre el icono del contraste cuyo resultado queremos conocer. En vez de hacer el proceso de selección que se mostró anteriormente, en «Codes for:» es posible especificar los dos grupos que se quieren comparar.



CUADRO IX.7. *(Continuación)*[STATISTICA]

1. *U* de Mann_Whitney

Mann-Whitney U Test (Cuadro IX.7.sta)
By variable Especie
Marked tests are significant at p <,05000
Include condition: v1="Cichla monoculus" or v1="Cichla orinocensis"

variable	Rank Sum Cichla monoculus	Rank Sum Cichla orinocensis	U	Z	p-level	Z adjusted	p-level	Valid N Cichla monoculus	Valid N Cichla orinocensis	2*1sided exact p
Tamaño ojo	294,0000	6,000000	0,00	2,749545	0,005968	2,749545	0,005968	21	3	0,000988

La tabla nos indica la suma de rangos de cada grupo, así como el estadístico *U*. Con esta *U* calcula dos valores *Z* (ajuste a la normal). El primero de ellos es el obtenido a partir de los datos brutos (Z = 2,7, p = 0,006) y el segundo considerando los rangos solapados ($Z_{adjusted}$ = 2,7, p = 0,006). En la última columna nos indica el valor *p* exacto (los otros dos son aproximados), que solo calcula cuando los tamaños de muestra son pequeños, y que no está corregido para considerar los empates (p < 0,001). En los tres casos p < 0,05 y, por tanto, se rechaza la hipótesis nula de homogeneidad, concluyendo que existen diferencias significativas en el tamaño del ojo en relación al tamaño del cuerpo, entre las dos especies.

2. Wald-Wolfowitz

Wald-Wolfowitz Runs Test (Cuadro IX.7.sta)
By variable Especie
Marked tests are significant at p <,05000
Include condition: v1="Cichla monoculus" or v1="Cichla orinocensis"

Variable	Valid N Cichla monoculus	Valid N Cichla orinocensis	Mean Cichla monoculus	Mean Cichla orinocensis	Z	p-level	Z adjstd	p-level	No. of Runs	No. of ties
Tamaño ojo	21	3	0,085314	0,064400	-4,31498	0,000016	3,807335	0,000140	2	0

La tabla nos indica las medias de cada grupo, así como dos estadígrafos de ajuste a la normal. El primero de ellos (Z = -4,31, p < 0,001) es para muestras grandes, mientras que $Z_{adjusted}$ (Z = 3,8, p < 0,001) incluye una corrección por continuidad, adecuada para muestras pequeñas. En nuestro caso el número de datos es pequeño y, por tanto, consideraríamos la segunda *Z,* aunque en este ejemplo no importa ya que ambas probabilidades son menores de 0,05 y, por tanto, se rechaza de nuevo la hipótesis nula de que las muestras son homogéneas.

3. Kolmogorov-Smirnov para dos muestras

Kolmogorov-Smirnov Test (Cuadro IX.7.sta)
By variable Especie
Marked tests are significant at p <,05000
Include condition: v1="Cichla monoculus" or v1="Cichla orinocensis"

variable	Max Neg Differnc	Max Pos Differnc	p-level	Mean Cichla monoculus	Mean Cichla orinocensis	Std.Dev. Cichla monoculus	Std.Dev. Cichla orinocensis	Valid N Cichla monoculus	Valid N Cichla orinocensis
Tamaño ojo	0,00	1,000000	p < .025	0,085314	0,064400	0,008145	0,000529	21	3

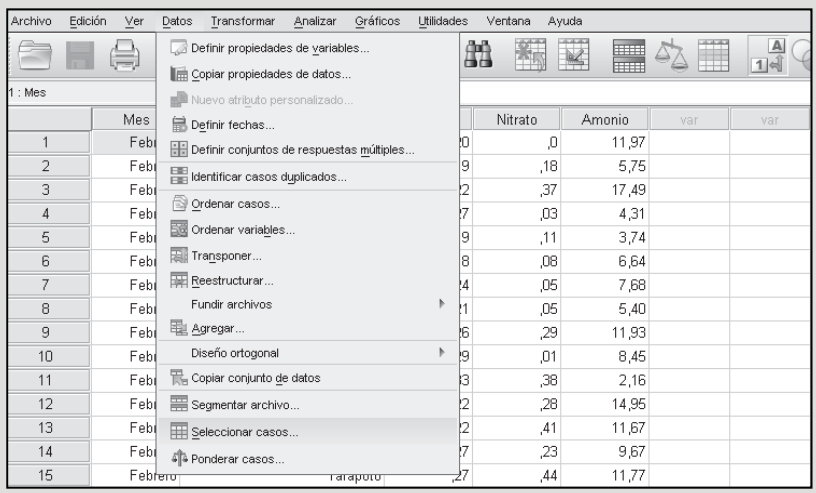
La tabla nos indica las diferencias máximas tanto negativas como positivas del análisis, así como el valor *p* del contraste. Como se puede comprobar p < 0,025 rechazando la hipótesis nula de que las muestras son homogéneas.

CUADRO IX.7. *(Continuación)*[SPSS]

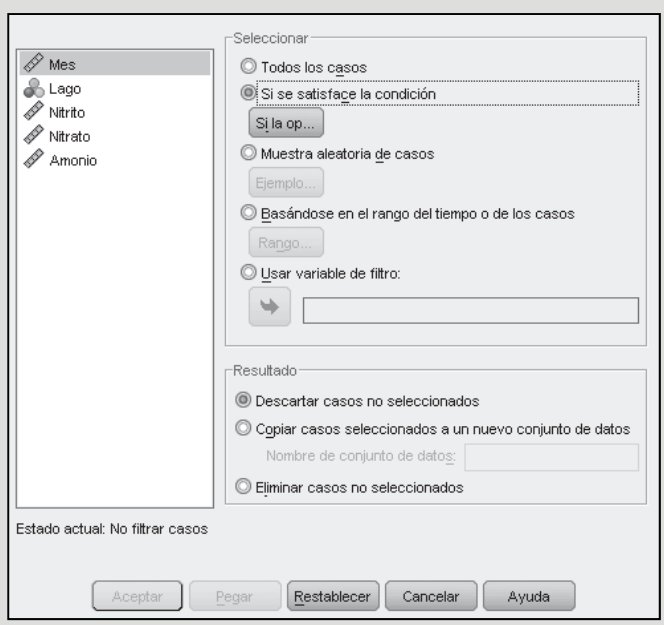
Aplicación con SPSS.

EJEMPLO. En el archivo **Cuadro IX.7.sav** hay datos de la concentración de nitrito, nitrato y amonio ($\mu M\ l^{-1}$) en lagos de la amazonia colombiana en dos meses diferentes. El objetivo es determinar si existen diferencias entre los lagos Correo y Tarapoto en el mes de febrero para la concentración de nitrato.

Paso 1. Como hay dos meses y solo se quiere trabajar con febrero, en primer lugar vamos a «Datos» ⇨ «Seleccionar casos…».

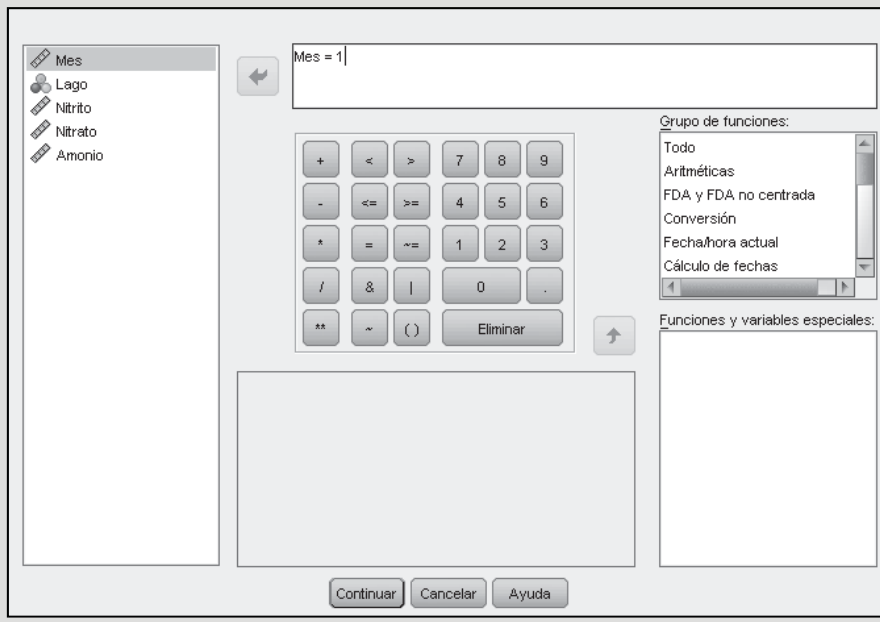


Paso 2. Se selecciona «Si se satisface la condición» y «Si la op…».

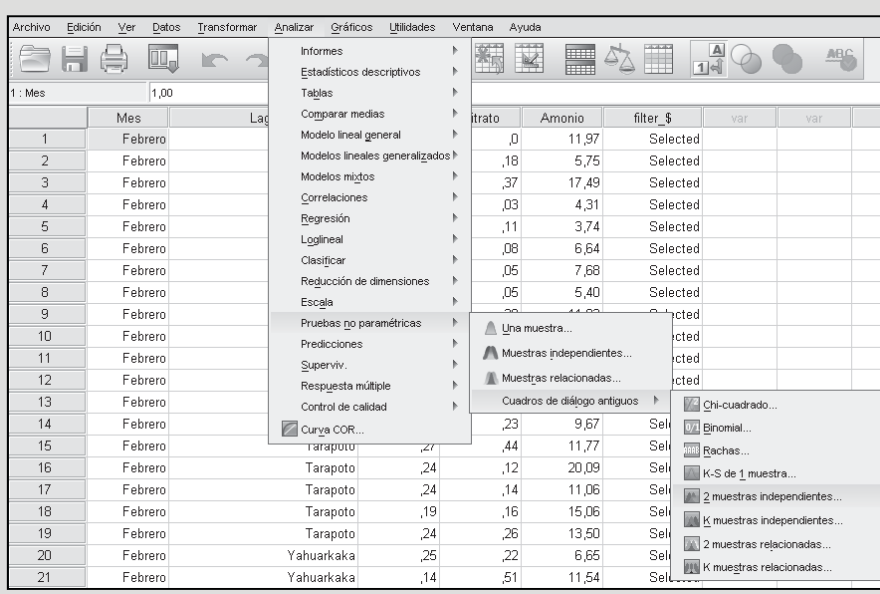


CUADRO IX.7. *(Continuación)*[SPSS]

Paso 3. Se escoge la variable, en este caso «Mes», y se pone la condición «Mes=1», que corresponde al mes de febrero. Después de pulsar «Continuar», cuando sale la ventana del paso 2 se pulsa «Aceptar».

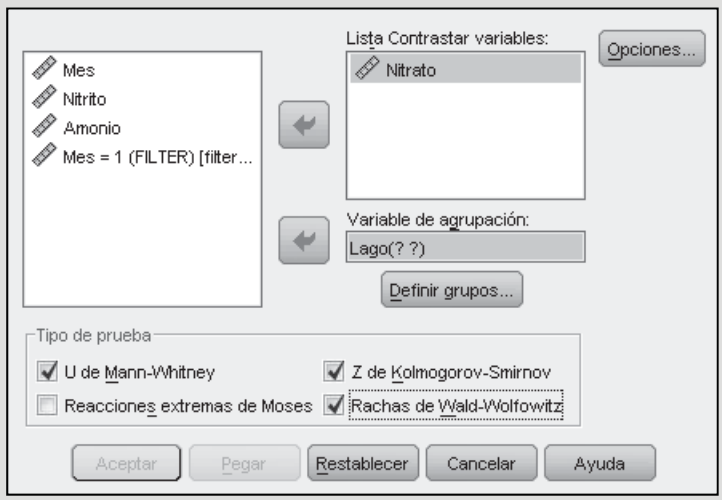


Paso 4. Se selecciona «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «2 muestras independientes…».

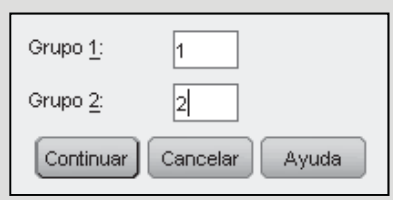


CUADRO IX.7. *(Continuación)*[SPSS]

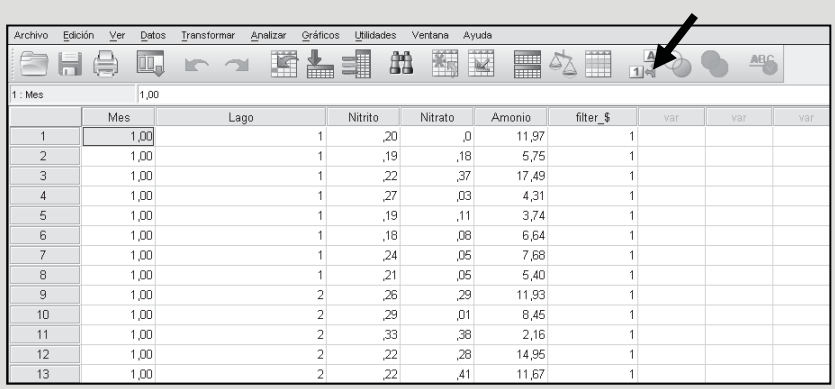
Paso 5. En «Lista Contrastar variables:» se introduce «Nitrato» y en «Variable de agrupación» se pone «Lago». En «Tipo de prueba» se seleccionan los test «U de Mann-Whitney», «Z de Kolmogorov-Smirnov» y «Rachas de Wald-Wolfowitz».



Paso 6. En la ventana anterior al pulsar en «Definir grupos» aparece la siguiente ventana donde se especifican los grupos a comparar, en este caso «1» y «2» que corresponden a los lagos Correo y Tarapoto.



Si existe alguna duda sobre los códigos de cada grupo, en la matriz de datos principal pulsando en el icono que muestra la flecha, aparecen los códigos de cada grupo en vez de los nombres.



	Mes	Lago	Nitrito	Nitrato	Amonio	filter_$	var	var	var
1	1,00	1	,20	,0	11,97	1			
2	1,00	1	,19	,18	5,75	1			
3	1,00	1	,22	,37	17,49	1			
4	1,00	1	,27	,03	4,31	1			
5	1,00	1	,19	,11	3,74	1			
6	1,00	1	,18	,08	6,64	1			
7	1,00	1	,24	,05	7,68	1			
8	1,00	1	,21	,05	5,40	1			
9	1,00	2	,26	,29	11,93	1			
10	1,00	2	,29	,01	8,45	1			
11	1,00	2	,33	,38	2,16	1			
12	1,00	2	,22	,28	14,95	1			
13	1,00	2	,22	,41	11,67	1			

CUADRO IX.7. *(Continuación)*[SPSS]

Paso 7. En el test de Mann-Whitney p = 0,026 y en el de Kolmogorov-Smirnov p = 0,036, por tanto, en ambos casos se rechaza la hipótesis nula de homogeneidad. Sin embargo, en la prueba de Wald-Wolfowitz p = 0,199 y se aceptaría la hipótesis nula. En resumen, de tres estadísticos posibles tenemos dos que rechazan la hipótesis nula y uno que la acepta. Sin embargo, el contraste de Wald-Wolfowitz es el menos sensible y, además, está pensado para la no existencia de rangos solapados. Por tanto, rechazamos la hipótesis nula de que las muestras son homogéneas y, concluimos que existen diferencias significativas en la concentración de nitrato entre los lagos Correo y Tarapoto.

Prueba de Mann-Whitney

Estadísticos de contraste[b]

	Nitrato
U de Mann-Whitney	17,000
W de Wilcoxon	53,000
Z	-2,229
Sig. asintót. (bilateral)	,026
Sig. exacta [2*(Sig. unilateral)]	,026[a]

a. No corregidos para los empates.

b. Variable de agrupación: Lago

Prueba de Kolmogorov-Smirnov para dos muestras

Estadísticos de contraste[a]

		Nitrato
Diferencias más extremas	Absoluta	,659
	Positiva	,000
	Negativa	-,659
Z de Kolmogorov-Smirnov		1,418
Sig. asintót. (bilateral)		,036

a. Variable de agrupación: Lago

Prueba de Wald-Wolfowitz

Estadísticos de contraste[b,c]

		Número de rachas	Z	Sig. exacta (unilateral)
Nitrato	Número exacto de rachas	8[a]	-,855	,199

a. No se han hallado empates intra-grupo.

b. Prueba de Wald-Wolfowitz

c. Variable de agrupación: Lago

IX.2.2. CONTRASTES PARA *K*-MUESTRAS INDEPENDIENTES

Los contrastes más utilizados en este caso son los de la mediana y la prueba de Kruskal-Wallis**.**

IX.2.2.1. Contraste de la mediana

Consiste en determinar el valor mediano de nuestras observaciones incluyendo todas las *k* muestras a analizar, y tipificar todos los valores como positivos si están por encima de la mediana y negativos si están por debajo.

Si las muestras fuesen homogéneas, en cada una de ellas aproximadamente la mitad de los valores deberían estar por encima de la mediana y la otra mitad por debajo. Ésta será la hipótesis nula (que todas las muestras proceden de la misma población).

Con el número de datos positivos y negativos de cada muestra, construimos una tabla de contingencia 2x*k*, a la que se aplica un contraste χ^2 de homogeneidad con la hipótesis nula indicada, es decir, en cada muestra la frecuencia esperada de positivos es igual a las frecuencia esperada de negativos.

Este contraste es poco sensible, por lo que es más preciso utilizar la prueba de Kruskal-Wallis. Sin embargo, en casos en los que los órdenes tienen límites artificiales este contraste es el más adecuado.

Los pasos para la obtención de este contraste aparecen en el Cuadro IX.8.

IX.2.2.2 ANOVA de Kruskal-Wallis

Es el contraste no paramétrico análogo al análisis de varianza, por lo que se denomina ANOVA, aunque el estadístico de contraste tiene distribución Chi-cuadrado. Es la prueba más utilizada para más de dos muestras independientes, siendo mucho más sensible que el test de la mediana.

A nivel teórico es una extensión del contraste *U* de Mann-Whitney para más de dos muestras. Esto quiere decir que es un test que mide la tendencia central de las muestras, teniendo como hipótesis nula que las poblaciones testadas tienen la misma mediana.

El desarrollo completo del cálculo de este estadístico se encuentra en Sokal & Rohlf (1981), en el que los pasos a seguir son:

1. Ordenar los datos en rangos (todas las muestras en conjunto).
2. Sumar los rangos de cada muestra y calcular el estadístico de contraste *H* que se compara con un valor de contraste tabulado.
3. Si *H* es mayor que el valor tabulado se rechaza la hipótesis nula, indicando que las muestras son diferentes.

En el Cuadro IX.8 se muestra un ejemplo de aplicación de esta prueba.

CUADRO IX.8. Contrastes de homogeneidad no paramétricos para *K* muestras independientes

Aplicación con R.

EJEMPLO. Continuando con el ejemplo del Cuadro IX.7, en el archivo **Cuadro IX.7.V.csv** hay datos de la concentración de nitrito, nitrato y amonio ($\mu M\ l^{-1}$) en lagos de la amazonia colombiana en dos meses diferentes. Ahora el objetivo es determinar si existen diferencias en la concentración de nutrientes considerando los tres lagos, para cada uno de los meses.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro IX.8.R**, que se muestra en la siguiente ventana. Sólo es necesario especificar lo que se solicita en la sección modificable por el usuario.

```
#Cuadro IX.8
#Contrastes de homogeneidad no paramétricos para K muestras independientes. Prueba de Kruskal-Wallis.

################################################
#Sección modificable por el usuario
################################################

#Lectura de datos.
datos<-read.csv2("Cuadro IX.7.V.csv",header=TRUE,encoding="latin1")

#Selección variable dependiente.
#varInteres<-c("Nitrito")
varInteres<-c("Nitrato")
#varInteres<-c("Amonio")

#Variable que corresponde al factor.
varFactor<-c("Lago")
#varFactor<-c("Mes")

##Variables de selección.
#Los elementos de las listas varSeleccion pueden modificarse en función de nuestros intereses.
#Pueden escogerse hasta dos grupos en cada variable de selección, en un número igual o inferior al n°
#varSeleccion1<-list(c("Lago","Correo","Tarapoto"))
varSeleccion1<-NULL
varSeleccion2<-list(c("Mes","Febrero"))
#varSeleccion2<-NULL
```

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro IX.7.V.csv**.

```
datos<-read.csv2("Cuadro IX.7.V.csv",header=TRUE,encoding="latin1")
```

Paso 2. Se selecciona la variable dependiente, en este caso se realiza el ejemplo para la concentración de nitratos.

```
#varInteres<-c("Nitrito")
varInteres<-c("Nitrato")
#varInteres<-c("Amonio")
```

Paso 3. A continuación se indica cual es el factor. En el ejemplo, como se quiere determinar las diferencias entre los lagos, este es el factor que se selecciona.

```
varFactor<-c("Lago")
#varFactor<-c("Mes")
```

CUADRO IX.8. *(Continuación)*[R]

Paso 4. También es posible seleccionar grupos en dos variables diferentes (un máximo de dos grupos en cada variable con este *script*). En el ejemplo, como se quiere comparar sólo el mes de febrero, se especifica en la variable selección y se pone el símbolo # delante del «NULL». Se observa que sólo se utiliza la variable selección 2, ya que en la 1 está activada la instrucción «NULL».

```
#varSeleccion1<-list(c("Lago","Correo","Tarapoto"))
varSeleccion1<-NULL
varSeleccion2<-list(c("Mes","Febrero"))
#varSeleccion2<-NULL
```

Paso 5. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
ArchivodeSalida<-"Salida Cuadro IX.8.txt"
```

Paso 6. El resultado del test tiene una p = 0,114, por lo que se acepta la hipótesis nula de homogeneidad, es decir, no existen diferencias significativas entre lagos en la concentración de nitratos en el mes de febrero.

```
[[1]]

        Kruskal-Wallis rank sum test

data:  datos2$valores and datos2$factores
Kruskal-Wallis chi-squared = 4.3487, df = 2, p-value = 0.1137
```

Se repite el proceso para el amonio (p = 0,205) y nitritos (p = 0,22), y en ambos casos también se acepta la hipótesis nula, no existen diferencias significativas entre los lagos en la concentración de estos nutrientes en el mes de febrero.

```
[[1]]

        Kruskal-Wallis rank sum test

data:  datos2$valores and datos2$factores
Kruskal-Wallis chi-squared = 3.1687, df = 2, p-value = 0.2051
```

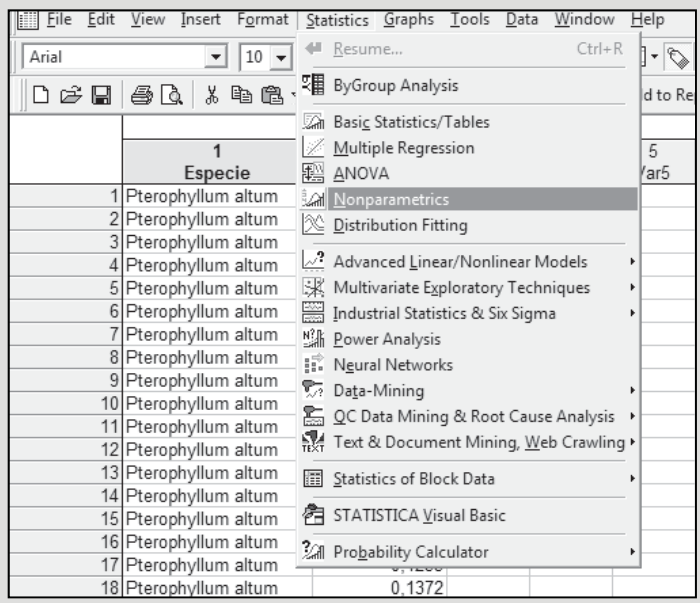
```
[[1]]

        Kruskal-Wallis rank sum test

data:  datos2$valores and datos2$factores
Kruskal-Wallis chi-squared = 3.0278, df = 2, p-value = 0.2200
```

CUADRO IX.8. *(Continuación)*[STATISTICA]

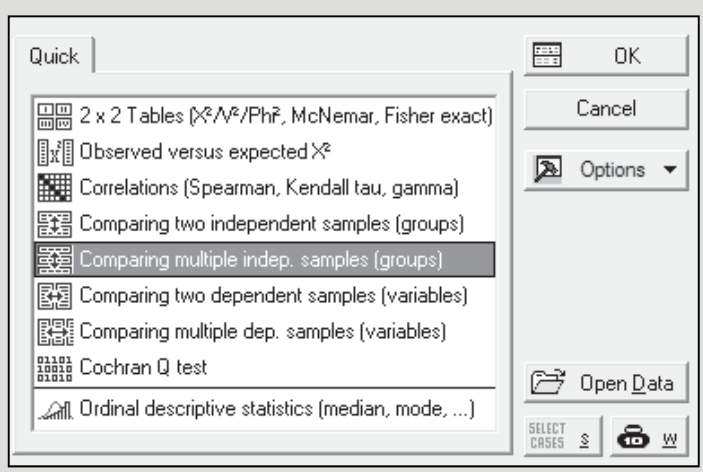
Aplicación con STATISTICA.

EJEMPLO. Continuando con el ejemplo del Cuadro IX.7, en el archivo **Cuadro IX.7.sta** hay datos de la relación tamaño del ojo/longitud del cuerpo de varias especies de perciformes de la familia Cichlidae. En este caso, el objetivo es determinar si existen diferencias entre todas las especies.

Paso 1. Se selecciona «Statistics» ⇨ «Nonparametrics».

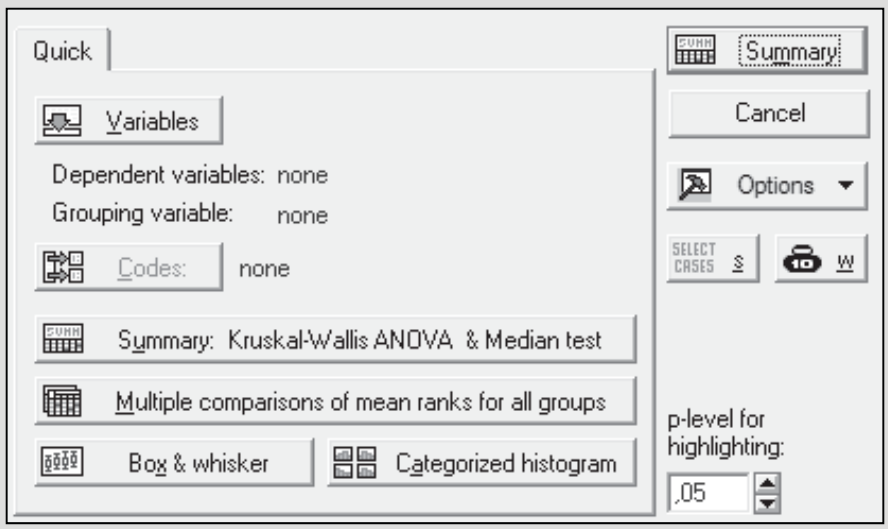


Paso 2. Seleccionamos el icono «Comparing multiple indep. samples (groups)».

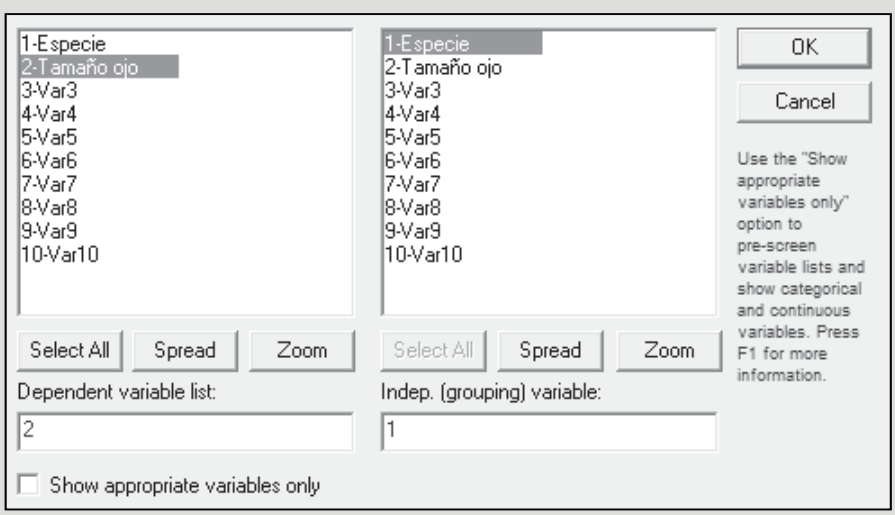


CUADRO IX.8. *(Continuación)* STATISTICA

Paso 3. Pulsamos el icono «Variables».

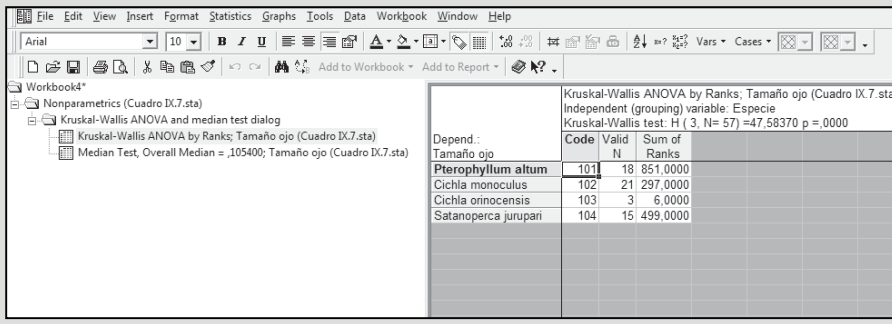


Paso 4. A la izquierda se indica la variable que contiene las mediciones «Tamaño ojo» y a la derecha la variable que indica el grupo de pertenencia «Especie».



Paso 5. En la ventana del paso 3 se pulsa sobre el icono «Summary: Kruskal-Wallis ANOVA & Median test», para que se muestren las tablas de resultados. Es importante indicar que aparecen dos tablas: una con el ANOVA de Kruskal-Wallis y otra con el test de la mediana. En la ventana de la izquierda se selecciona el test que se desea ver.

CUADRO IX.8. *(Continuación)* STATISTICA


File Edit View Insert Format Statistics Graphs Tools Data Workbook Window Help
Workbook4*
Nonparametrics (Cuadro IX.7.sta)
Kruskal-Wallis ANOVA and median test dialog
Kruskal-Wallis ANOVA by Ranks; Tamaño ojo (Cuadro IX.7.sta)
Median Test, Overall Median = ,105400; Tamaño ojo (Cuadro IX.7.sta)
Kruskal-Wallis ANOVA by Ranks; Tamaño ojo (Cuadro IX.7.sta)
Independent (grouping) variable: Especie
Kruskal-Wallis test: H (3, N= 57) =47,58370 p =,0000
Depend.: Tamaño ojo
Code
Valid N
Sum of Ranks
Pterophyllum altum 101 18 851,0000
Cichla monoculus 102 21 297,0000
Cichla orinocensis 103 3 6,0000
Satanoperca jurupari 104 15 499,0000

1. ANOVA de Kruskal-Wallis

	Kruskal-Wallis ANOVA by Ranks; Tamaño ojo (Cuadro IX.7.sta) Independent (grouping) variable: Especie Kruskal-Wallis test: H (3, N= 57) =47,58370 p =,0000		
Depend.: Tamaño ojo	Code	Valid N	Sum of Ranks
Pterophyllum altum	101	18	851,0000
Cichla monoculus	102	21	297,0000
Cichla orinocensis	103	3	6,0000
Satanoperca jurupari	104	15	499,0000

La tabla nos devuelve el valor de H = 47,6 para 3 grados de libertad y 57 datos, teniendo asociada un probabilidad $p < 0{,}001$, por lo que se rechaza la hipótesis nula de homogeneidad.

2. Prueba de la Mediana

	Median Test, Overall Median = ,105400; Tamaño ojo (Cuadro IX.7.sta) Independent (grouping) variable: Especie Chi-Square = 43,66256 df = 3 p = ,0000				
Dependent: Tamaño ojo	Pterophyllum altum	Cichla monoculus	Cichla orinocensis	Satanoperca jurupari	Total
<= Median: observed	0,00000	21,0000	3,00000	5,00000	29,00000
expected	9,15789	10,6842	1,52632	7,63158	
obs.-exp.	-9,15789	10,3158	1,47368	-2,63158	
> Median: observed	18,00000	0,0000	0,00000	10,00000	28,00000
expected	8,84211	10,3158	1,47368	7,36842	
obs.-exp.	9,15789	-10,3158	-1,47368	2,63158	
Total: observed	18,00000	21,0000	3,00000	15,00000	57,00000

Nos indica el valor de la prueba Chi-cuadrado realizada, donde χ^2 = 43,6 con 3 grados de libertad y una probabilidad $p < 0{,}001$. Con este resultado se rechaza la hipótesis nula de que las muestras son homogéneas.

Por tanto, ambas pruebas rechazan la hipótesis nula, por lo que se puede concluir que existen diferencias significativas en la relación tamaño del ojo/longitud del cuerpo entre las diferentes especies de peces.

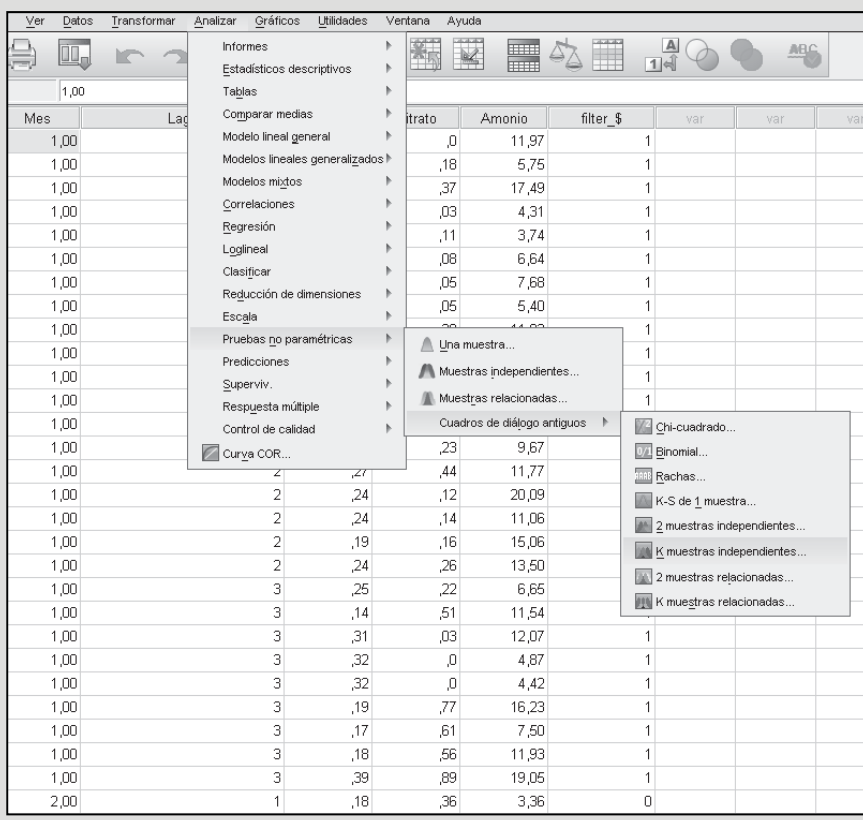
CUADRO IX.8. *(Continuación)*[SPSS]

Aplicación con SPSS.

EJEMPLO. Continuando con el ejemplo del Cuadro IX.7, en el archivo **Cuadro IX.7.sav** hay datos de la concentración de nitrito, nitrato y amonio ($\mu M\ l^{-1}$) en lagos de la amazonia colombiana en dos meses diferentes. Ahora el objetivo es determinar si existen diferencias entre todos los lagos en el mes de febrero para la concentración de nitrato.

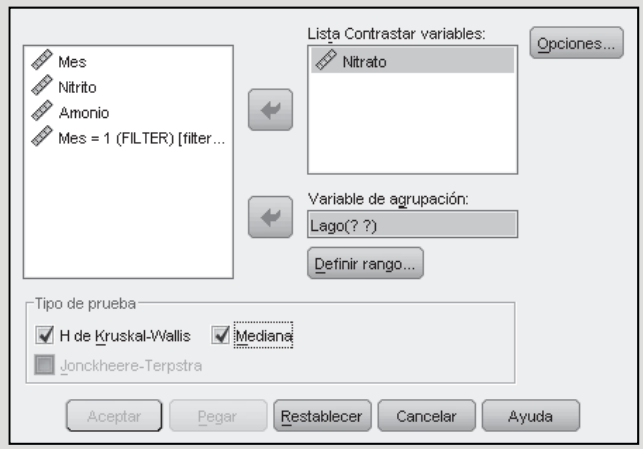
Paso 1. Como hay dos meses y solo se quiere trabajar con febrero, se hace la selección de los casos como se explicó en los pasos 1 a 3 del Cuadro IX.7.

Paso 2. Se selecciona «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «k muestras independientes…».

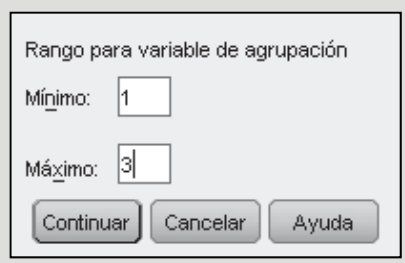


Paso 3. En «Lista Contrastar variables:» se introduce «Nitrato» y en «Variable de agrupación» se pone «Lago». En «Tipo de prueba» se seleccionan los test «H de Kruskal-Wallis» y «Mediana».

CUADRO IX.8. *(Continuación)*[SPSS]



Paso 4. En la ventana anterior al pulsar en «Definir rango…» aparece la siguiente ventana donde se especifican los grupos a comparar, en este caso desde «1» a «3» que corresponden a los tres lagos.



Paso 5. El test de Kruskal-Wallis acepta la hipótesis nula de homogeneidad (p = 0,107), pero el test de la mediana lo rechaza (p = 0,042), aunque está muy cerca el nivel de significación de 0,05. En este caso la prueba de la mediana no es aplicable, dado que –como nos advierten los resultados– el 66,7% de las casillas tienen frecuencias esperadas menores que 5, y ese porcentaje no debe superar el 20%. Por lo tanto, interpretamos solamente el test de Kruskal-Wallis, y concluimos que no existen diferencias significativas en la concentración de nitratos entre los tres lagos en el mes de febrero.

Prueba de Kruskal-Wallis

Estadísticos de contraste[a,b]

	Nitrato
Chi-cuadrado	4,462
gl	2
Sig. asintót.	,107

Prueba de la mediana

Estadísticos de contraste[b]

	Nitrato
N	28
Mediana	,1985
Chi-cuadrado	6,318[a]
gl	2
Sig. asintót.	,042

IX.2.3. CONTRASTES PARA DOS MUESTRAS DEPENDIENTES

Las pruebas más utilizadas son la del contraste de los signos y la prueba de Wilcoxon para pares relacionados.

IX.2.3.1. Contraste de los signos

Este contraste mide la dirección de las diferencias entre dos muestras pareadas, por lo que es muy útil cuando se trabaja con muestras pareadas a las que se les puede identificar un incremento o descenso en la medición, pero es difícil de cuantificar este cambio. En otras condiciones, el test de Wilcoxon es más sensible.

Se estudia el número de veces que las diferencias de los valores de las dos muestras son positivas, negativas o si no hay diferencias. Con estas premisas podemos contrastar la hipótesis nula de que las muestras son homogéneas, ya que en ese caso habrá aproximadamente tantas diferencias negativas como positivas.

El desarrollo completo de la prueba aparece en Siegel & Castellan (1988), pero los pasos de forma resumida son:

1. Determinación de las diferencias (+ ó -).
2. Ajuste de las diferencias observadas a una distribución binomial. Cuando el número de datos es grande (mayor de 30 datos), se utiliza una aproximación a la distribución Normal (cálculo del estadístico *Z*) de la distribución binomial.
3. Determinación de la probabilidad asociada al estadístico de contraste para aceptar o rechazar la hipótesis nula de que las muestras son homogéneas.

Los pasos para realizar este contraste aparecen en el Cuadro IX.9.

IX.2.3.2. Prueba de Wilcoxon para pares relacionados

Es la prueba análoga a la *t* de Student para muestras relacionadas, siendo casi tan potente como ésta. En lo que se refiere a la sensibilidad, es un contraste mucho más potente que el anterior ya que, aunque también utiliza las diferencias entre los valores de cada caso, el valor absoluto de las diferencias se ordenan en rangos (no simplemente positivo y negativo), habiendo más información sobre las diferencias que en el caso anterior.

Si las muestras son homogéneas, la hipótesis nula, la suma de los rangos de las diferencias positivas tiene que ser similar a la suma de los rangos con valor negativo (aleatoriedad de las diferencias).

Una descripción detallada de esta prueba se encuentra en Siegel & Castellan (1988), los pasos principales son:

1. Ordenar los valores de las muestras de forma ascendente y ordenar en rangos las diferencias en valor absoluto entre ambas muestras,
2. Calcular el estadístico de contraste *T* y *T'.*

3. Si *T* ó T′ es menor o igual que las cantidades límite que aparecen en la denominada tabla de Wilcoxon, se rechaza la hipótesis nula de que las variables son homogéneas.

En caso de que el tamaño de muestra sea alto (mayor de 100 datos), se puede hacer una aproximación a la Normal del estadístico *T*, pudiéndose calcular el valor *Z* de la distribución Normal que nos devuelve la probabilidad del contraste.

En el Cuadro IX.9 se describe el procedimiento para la obtención de este contraste.

CUADRO IX.9. Contrastes de homogeneidad no paramétricos para dos muestras dependientes

Aplicación con R.

EJEMPLO. En el archivo **Cuadro IX.9.V.csv** hay datos de la concentración de eritrocitos (en millones por milímetro cúbico), de varios hombres y mujeres que fueron sometidos a un tratamiento para aumentar la concentración de glóbulos rojos. El objetivo es determinar si la concentración es diferente entre el inicio y al cabo de un mes, haciendo la diferenciación entre hombres y mujeres.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro IX.9.R**, que se muestra en la siguiente ventana. Sólo es necesario especificar lo que se solicita en la sección modificable por el usuario.

```
#Cuadro IX.9
#Contrastes de homogeneidad no paramétricos para 2 muestras dependientes. Test de Wilcoxon.

################################################
#Sección modificable por el usuario
################################################

#Lectura de datos.
datos<-read.csv2("Cuadro IX.9.V.csv",header=TRUE,encoding="latin1")

#Selección de las dos variables a comparar.
varInteres<-c("Inicio","Mes1")

##Variables de selección.
#Los elementos de las listas varSeleccion pueden modificarse en función de nuestros intereses.
#Pueden escogerse hasta dos grupos en cada variable de selección, en un número igual o inferio
varSeleccion1<-list(c("Sexo","Mujer"))
#varSeleccion1<-NULL
#varSeleccion2<-list(c("","",""))
varSeleccion2<-NULL

#Archivo de salida.
ArchivodeSalida<-"Salida Cuadro IX.9.txt"
```

Paso 1. Hay que definir el nombre del archivo desde donde se van a importar los datos, en este caso **Cuadro IX.9.V.csv**.

datos<-read.csv2("Cuadro IX.9.V.csv",header=TRUE,encoding="latin1")

CUADRO IX.9. *(Continuación)*[R]

Paso 2. A continuación hay que especificar las 2 variables que se comparan, en este caso la concentración de eritrocitos al inicio y al cabo de un mes.

```
varInteres<-c("Inicio","Mes1")
```

Paso 3. También es posible seleccionar grupos en dos variables diferentes (un máximo de dos grupos en cada variable con este *script*). En primer lugar comparamos sólo las mujeres.

```
varSeleccion1<-list(c("Sexo","Mujer"))
#varSeleccion1<-NULL
#varSeleccion2<-list(c("","",""))
varSeleccion2<-NULL
```

Paso 4. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
ArchivodeSalida<-"Salida Cuadro IX.9.txt"
```

Paso 5. El resultado de la prueba de Wilcoxon tiene una p = 0,959, por lo que se acepta la hipótesis nula de homogeneidad, es decir, no existen diferencias significativas en la concentración de eritrocitos entre el inicio y al cabo de un mes en las mujeres. Es importante mencionar que cuando existen empates, es decir, existen valores iguales en los grupos a comparar, R no calcula el valor exacto de probabilidad y, por ello, se pueden observar pequeñas diferencias con el resultado de SPSS y STATISTICA.

```
[[1]]

        Wilcoxon signed rank test with continuity correction

data:  datos2$var1 and datos2$var2
V = 26.5, p-value = 0.9592
alternative hypothesis: true location shift is not equal to 0
```

Se repite el proceso para los hombres y tampoco se observan diferencias significativas (p = 0,724).

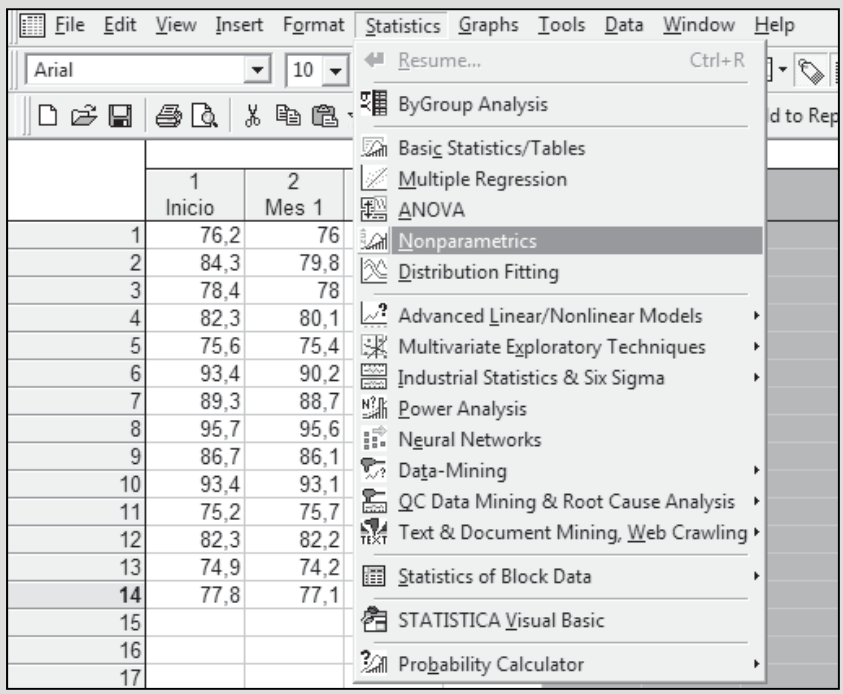
```
[[1]]

        Wilcoxon signed rank test with continuity correction

data:  datos2$var1 and datos2$var2
V = 34, p-value = 0.724
alternative hypothesis: true location shift is not equal to 0
```

CUADRO IX.9. *(Continuación)*[STATISTICA]

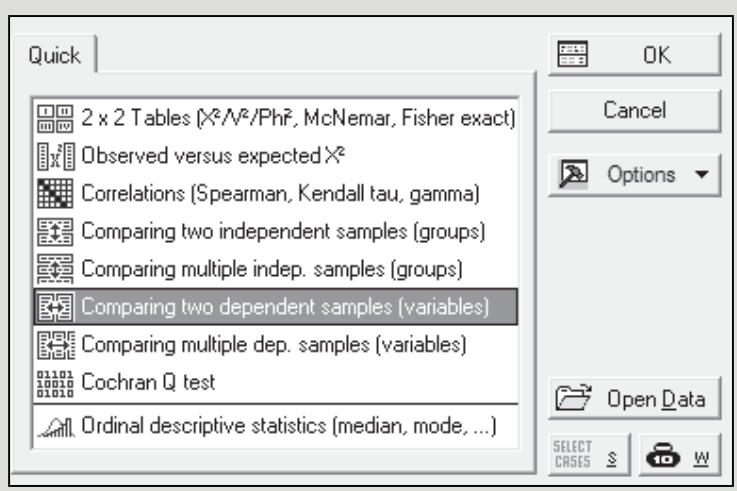
Aplicación con STATISTICA.

EJEMPLO. Se somete un grupo de 14 personas a un tratamiento de adelgazamiento durante tres meses. El objetivo es determinar si hay diferencias en el peso entre el inicio y el primer mes de tratamiento. Los datos se encuentran en el archivo **Cuadro IX.9.sta.**

Paso 1. Se selecciona «Statistics» ⇨ «Nonparametrics».

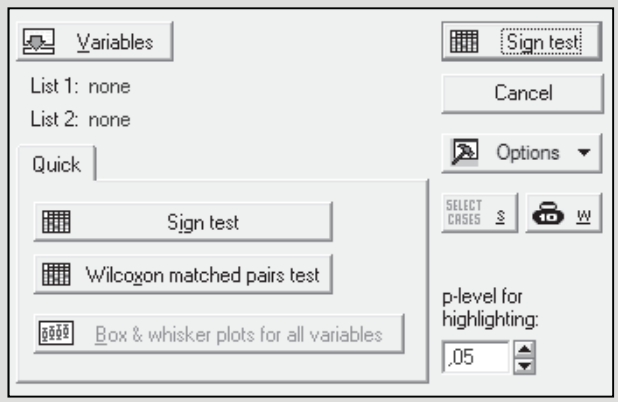


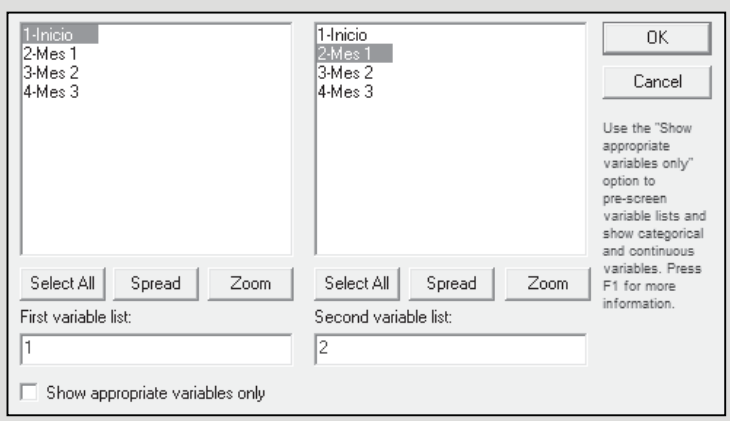
Paso 2. Seleccionamos «Comparing two dependent samples (variables)».



CUADRO IX.9. *(Continuación)* *STATISTICA*

Paso 3. Se pulsa en «Variables» y aparece la siguiente ventana donde se seleccionan las variables que queremos comparar.





Paso 4. En la ventana anterior pulsando sobre los iconos «Sign Test» y «Wilcoxon matched pair test» aparecen las tablas de resultados.

	Sign Test (Cuadro IX.9.sta) Marked tests are significant at p <,05000			
Pair of Variables	No. of Non-ties	Percent v < V	Z	p-level
Inicio & Mes 1	14	7,142857	2,939874	0,003283

	Wilcoxon Matched Pairs Test (Cuadro IX.9.sta) Marked tests are significant at p <,05000			
Pair of Variables	Valid N	T	Z	p-level
Inicio & Mes 1	14	7,000000	2,856330	0,004286

En ambas pruebas $p < 0,05$, por tanto existen diferencias significativas en el peso de las personas entre el inicio y el primer mes. Sin embargo, estas diferencias no significan que el primer mes el peso medio de las personas sea menor, simplemente que es diferente del peso medio al inicio.

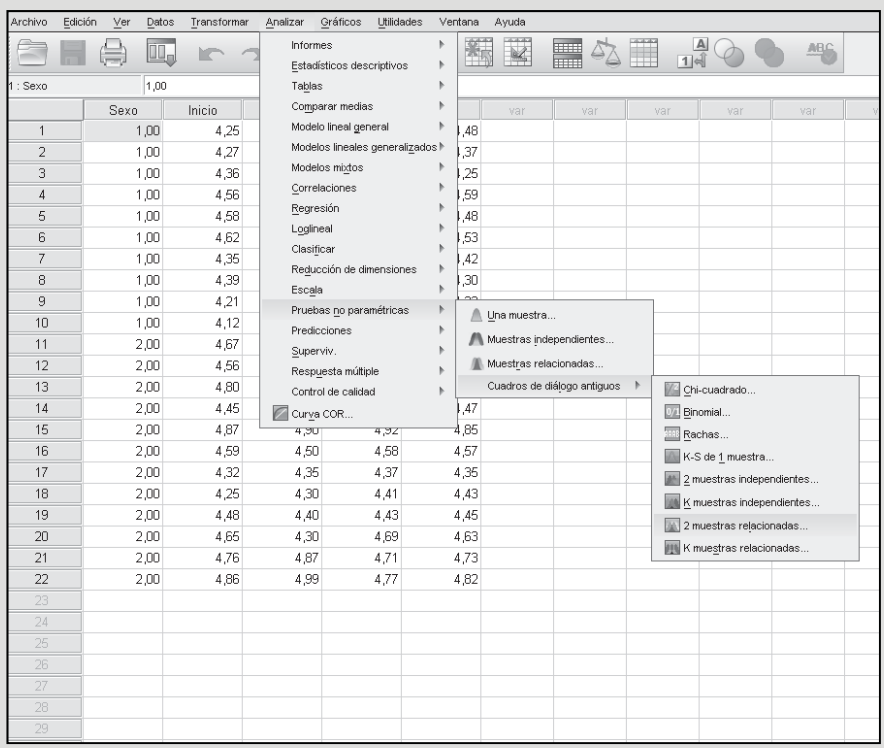
CUADRO IX.9. *(Continuación)*[SPSS]

Aplicación con SPSS.

EJEMPLO. Se usarán los mismos datos que en la aplicación con R, que son valores de la concentración de eritrocitos (en millones por milímetro cúbico), de varios hombres y mujeres que fueron sometidos a un tratamiento para aumentar la concentración de glóbulos rojos. El objetivo es determinar si la concentración es diferente entre el inicio y al cabo de un mes en los hombres. Los datos están en el archivo **Cuadro IX.9.sav**

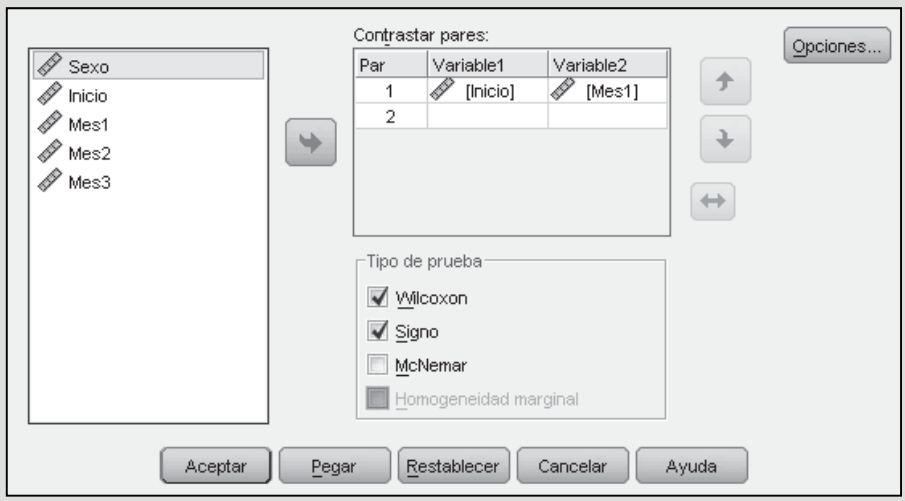
Paso 1. Como hay hombres y mujeres, y solo se quiere trabajar con el grupo de los hombres, se hace la selección de los casos como se explicó en los pasos 1 a 3 del Cuadro IX.7.

Paso 2. Se selecciona «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «2 muestras relacionadas…».



Paso 3. En «Contrastar pares:» en «Variable1» se pone «Inicio» y en «Variable2» se introduce «Mes1». En «Tipo de prueba» se selecciona «Wilcoxon» y «Signo». Es posible contrastar varios pares de variables a la vez, simplemente introduciendo un par 2, un par 3, etc.

CUADRO IX.9. *(Continuación)*[SPSS]



Paso 4. Los resultados muestran que en ambas pruebas se acepta la hipótesis nula de homogeneidad ya que la $p > 0,05$; Wilcoxon ($p = 0,695$) y contraste de los signos ($p = 0,774$). Por lo tanto, se concluye que no hay diferencias significativas en la concentración de glóbulos rojos en los hombres entre el inicio y el primer mes del estudio.

Prueba de los rangos con signo de Wilcoxon

Estadísticos de contraste[b]

	Mes1 - Inicio
Z	-,393[a]
Sig. asintót. (bilateral)	,695

a. Basado en los rangos negativos.

b. Prueba de los rangos con signo de Wilcoxon

Prueba de los signos

Estadísticos de contraste[b]

	Mes1 - Inicio
Sig. exacta (bilateral)	,774[a]

a. Se ha usado la distribución binomial.

b. Prueba de los signos

IX.2.4. CONTRASTES PARA *K*-MUESTRAS DEPENDIENTES

La prueba más utilizada es el ANOVA de Friedman. Esta prueba contrasta la hipótesis de que las medidas proceden de poblaciones iguales, cuando los datos de cada una de las muestras han sido transformados en rangos de forma ascendente.

Es la prueba análoga al ANOVA para muestras repetidas por lo que, a pesar de que se trata de un contraste χ^2, recibe el nombre de ANOVA.

Los pasos de esta prueba están descritos por Siegel & Castellan (1988). De forma resumida son:

1. Asignar rangos dentro de cada caso de la muestra.
2. Sumar los rangos de cada variable para toda la muestra.
3. Calcular el estadístico χ^2 de contraste y compararlo con el valor tabulado χ^2 crítico con *a* - 1 grados de libertad, siendo *a* el número de valores medidos para cada muestra.

Un ejemplo de la aplicación de este test aparece en el Cuadro IX.10.

CUADRO IX.10. Contrastes de homogeneidad no paramétricos para *k* muestras dependientes

Aplicación con R.

EJEMPLO. Continuando con el ejemplo del Cuadro IX.9, el objetivo ahora es determinar si la concentración de glóbulos rojos es diferente considerando todos los tiempos analizados, tanto en hombres como en mujeres. Los datos están en el archivo **Cuadro IX.10.V.csv**.

Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, abrimos el archivo **Cuadro IX.10.R**, que se muestra en la siguiente ventana. Sólo es necesario especificar lo que se solicita en la sección modificable por el usuario.

```
#Cuadro IX.10
#Contrastes de homogeneidad no paramétricos para k muestras dependientes. Test de Friedman.

################################################
#Sección modificable por el usuario
################################################

#Lectura de datos.
datos<-read.csv2("Cuadro IX.10.V.csv",header=TRUE,encoding="latin1")

#Selección de los grupos de la variable dependiente. Deben ser más de dos y hasta un máximo de 20.
varInteres<-c("Inicio","Mes1","Mes2","Mes3")

#Bloques
varBloque<-c("Individuo")

#Archivo de salida.
ArchivodeSalida<-"Salida Cuadro IX.10.txt"
```

CUADRO IX.10. *(Continuación)*[R]

Paso 1. En el archivo de datos cada caso tiene que ir identificado con una etiqueta distintiva, por ejemplo, «Ind1», «Ind2» ó «1», «2», «3», etc.

Individuo	Inicio	Mes1	Mes2	Mes3
Ind1	4,25	4,26	4,4	4,48
Ind2	4,27	4,25	4,3	4,37
Ind3	4,36	4,35	4,25	4,25
Ind4	4,56	4,6	4,5	4,59
Ind5	4,58	4,55	4,5	4,48
Ind6	4,62	4,6	4,45	4,53
Ind7	4,35	4,36	4,4	4,42
Ind8	4,39	4,4	4,32	4,3
Ind9	4,21	4,2	4,25	4,32
Ind10	4,12	4,2	4,15	4,14
Ind11	4,67	4,87	4,6	4,52
Ind12	4,56	4,5	4,57	4,55
Ind13	4,8	4,78	4,74	4,81
Ind14	4,45	4,5	4,43	4,47

Paso 2. En el *script*, en primer lugar hay que especificar cuáles son las variables dependientes, que deben ser más de dos y hasta un máximo de 20.

```
varInteres<-c("Inicio","Mes1","Mes2","Mes3")
```

Paso 3. A continuación se especifica la variable que identifica cada uno de los casos.

```
varBloque<-c("Individuo")
```

Paso 4. Por último, se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto.

```
ArchivodeSalida<-"Salida Cuadro IX.10.txt"
```

Los resultados del contraste de Friedman muestran que, en el conjunto de hombres y mujeres, la concentración de glóbulos rojos no cambia en el tiempo (p = 0,967).

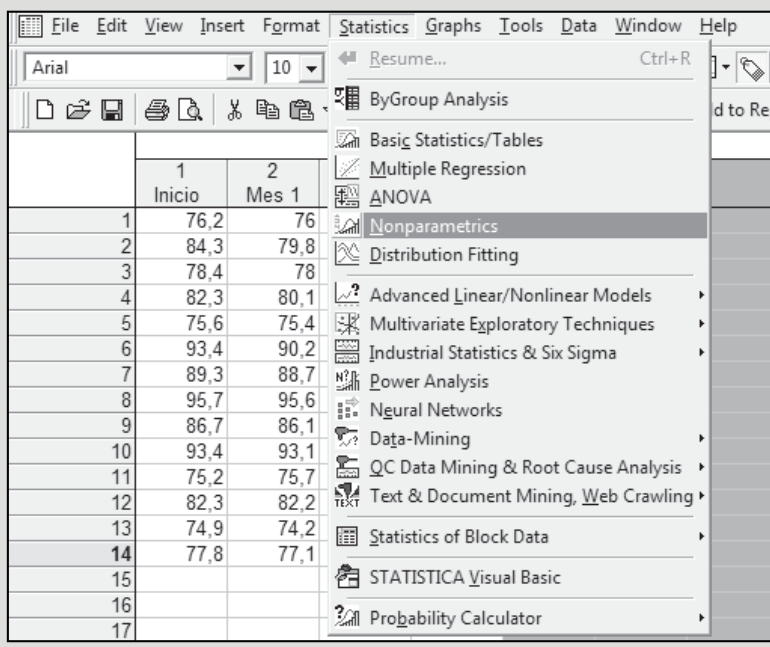
```
[[1]]

        Friedman rank sum test

data:  datos1$valores, datos1$variablestot and datos1$bloquetot
Friedman chi-squared = 0.2615, df = 3, p-value = 0.9671
```

CUADRO IX.10. *(Continuación)*[STATISTICA]

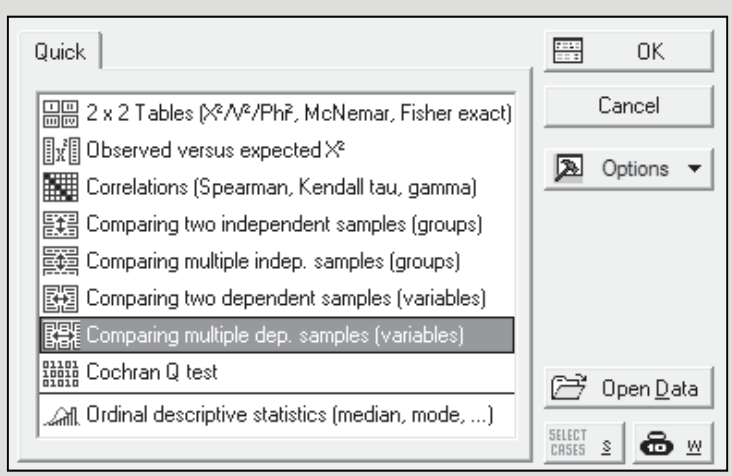
Aplicación con STATISTICA.

EJEMPLO. Continuando con el ejemplo del Cuadro IX.9 en el que un grupo de 14 personas seguía un régimen de adelgazamiento, nos planteamos ahora comprobar si hay diferencias en el peso medio entre todos los meses, incluyendo el peso inicial. Los datos se encuentran en el archivo **Cuadro IX.9.sta.**

Paso 1. Se selecciona «Statistics» ⇨ «Nonparametrics».

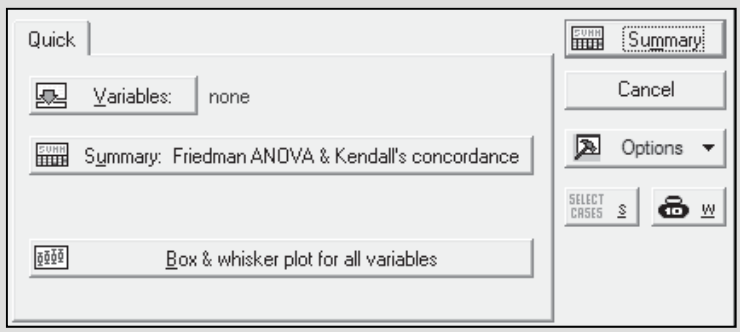


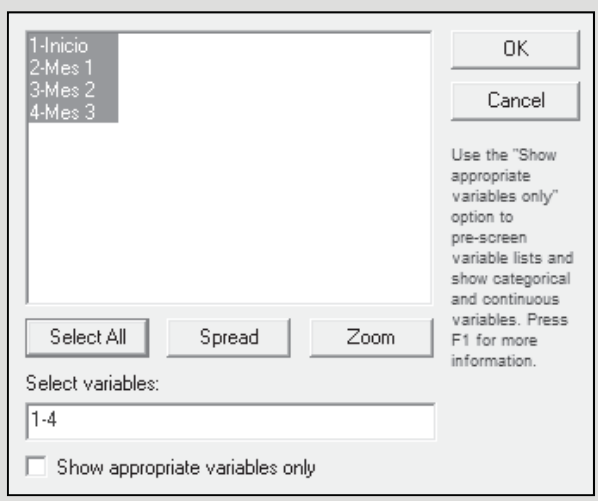
Paso 2. Seleccionamos «Comparing multiple dep. samples (variables)».



CUADRO IX.10. *(Continuación)* STATISTICA

Paso 3. Se pulsa en «Variables» y aparece la siguiente ventana donde se seleccionan todas las variables.





Paso 4. Pulsando sobre el icono «Summary: Friedman ANOVA & Kendall's concordance» aparece la tabla de resultados.

	Friedman ANOVA and Kendall Coeff. of Concordance (Cuadro IX.9.sta) ANOVA Chi Sqr. (N = 14, df = 3) = 38,65714 p = ,00000 Coeff. of Concordance = ,92041 Aver. rank r = ,91429			
Variable	Average Rank	Sum of Ranks	Mean	Std.Dev.
Inicio	3,928571	55,00000	83,25000	7,382073
Mes 1	2,928571	41,00000	82,30000	7,149180
Mes 2	2,142857	30,00000	81,87857	7,191574
Mes 3	1,000000	14,00000	81,10000	7,355898

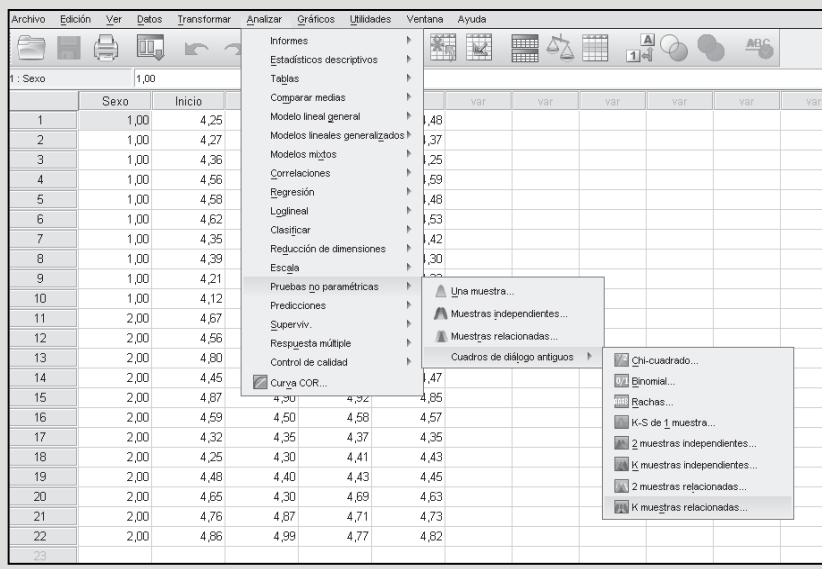
Como puede observarse, el contraste Chi-cuadrado indica que las muestras no son homogéneas (p < 0,001), es decir, el peso medio mensual de todas las personas es diferente al comparar todos los meses. Además, existe una alta asociación entre las variables (peso y tiempo), ya que el coeficiente de correlación de Spearman promedio es de 0,91 y el coeficiente de concordancia de 0,92.

CUADRO IX.10. *(Continuación)*[SPSS]

Aplicación con SPSS.

EJEMPLO. Continuando con el ejemplo del Cuadro IX.9, el objetivo ahora es determinar si la concentración de glóbulos rojos es diferente considerando todos los meses y el valor inicial, no haciendo la diferenciación entre hombres y mujeres. Los datos están en el archivo **Cuadro IX.9.sav**

Paso 1. Se selecciona «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «k muestras relacionadas…».



Paso 2. En «Variables de contraste» se incluyen todos los tiempos analizados y se selecciona «Friedman» y «W de Kendall».

CUADRO IX.10. *(Continuación)*[SPSS]

Paso 3. El contraste de Friedman acepta la hipótesis nula de que las muestras son homogéneas (p = 0,967), es decir, la concentración de glóbulos rojos no cambia en el tiempo en el conjunto de hombres y mujeres. Además, existe una muy baja asociación entre las variables (concentración de glóbulos rojos y tiempo), ya que el coeficiente de concordancia es 0,004.

Prueba de Friedman

Estadísticos de contraste[a]

N	22
Chi-cuadrado	,261
gl	3
Sig. asintót.	,967

a. Prueba de Friedman

Prueba W de Kendall

Estadísticos de contraste

N	22
W de Kendall[a]	,004
Chi-cuadrado	,261
gl	3
Sig. asintót.	,967

a. Coeficiente de concordancia de Kendall

Capítulo X

CONTRASTES DE INDEPENDENCIA Y ASOCIACIÓN EN VARIABLES CUANTITATIVAS

Uno de los objetivos más frecuentes a la hora de analizar datos es el de estudiar el grado de dependencia o asociación entre variables. La *correlación* es el marco teórico general en el que podremos estudiar estos aspectos.

Decimos que una variable *y* depende de otra variable *x* cuando el valor de *x* determina en cierta medida el valor de *y*. Puede darse el caso de que el valor de *y* quede perfectamente determinado por el valor de *x*. Entonces, si conocemos el valor de la variable *x*, hay una relación funcional que nos define de forma perfecta el valor que tomaría *y*. En este caso hablaríamos de que entre ambas variables existe una relación de *dependencia funcional*. Pero en la práctica, con mayor frecuencia nos encontraremos con situaciones en las que el valor de la variable *y* no queda perfectamente definido por el valor de *x*, sino que existe un cierto grado de relación entre las dos que nos permite aproximar el valor de *y*. En este caso hablamos de una relación entre ambas variables de *dependencia aleatoria*.

Para analizar el grado de dependencia o asociación entre variables utilizamos una serie de estadísticos que se denominan *coeficientes de correlación*. Los coeficientes de correlación nos indican si las variables están asociadas o no (si son independientes) y además, en qué sentido se da esa asociación o correlación entre variables (positiva o negativa). También nos informan acerca de la intensidad de la correlación. A través de diversos estadísticos, también podemos saber si la correlación entre variables es significativa o no.

X.1. Correlaciones bivariadas simples

X.1.1. COEFICIENTE DE CORRELACIÓN DE PEARSON (*r*)

El coeficiente de correlación de Pearson (*r*) indica la fuerza de la relación linea entre las dos variables. Es necesario asumir, por tanto, que cualquier relación de dependencia que exista entre ambas variables es de tipo lineal. Solo sirve para variables cuantitativas.

El coeficiente *r* de Pearson se calcula según la siguiente fórmula, que se basa en la covarianza entre las dos variables:

$$r_{y_1,y_2} = \frac{\sum_{i=1}^{n}\left[\left(y_{i1}-\bar{y}_1\right)\cdot\left(y_{i2}-\bar{y}_2\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(y_{i1}-\bar{y}_1\right)^2\sum_{i=1}^{n}\left(y_{i2}-\bar{y}_2\right)^2}}$$

o lo que es lo mismo:

$$\frac{S_{xy}}{S_x S_y}$$

es decir, la covarianza dividida por el producto de las desviaciones típicas.

Los valores de este coeficiente se extienden en un rango entre -1 y 1. Cuanto más cerca de -1 o 1, más fuerte es la correlación entre las variables. Valores cercanos a 0 indican ausencia de correlación y, por tanto, de dependencia.

Para la prueba de significación del coeficiente de Pearson suele establecerse como H_0 que $R = 0$, utilizando el coeficiente de correlación r obtenido en la muestra como estimador del verdadero coeficiente de correlación poblacional R. Además es necesario para esta prueba que la distribución de probabilidad conjunta de ambas variables sea Normal.

$$t = \frac{r}{s_r}$$

donde s_r es el error estándar del estimador r:

$$s_r = \sqrt{\frac{(1 - r^2)}{(n - 2)}}$$

En el Cuadro X.1 mostramos un ejemplo de cálculo del coeficiente de correlación de Pearson.

X.1.2. COEFICIENTE DE CORRELACIÓN DE SPEARMAN (ρ)

Entre los métodos no paramétricos más utilizados nos encontramos con el coeficiente de Spearman (ρ). Como en todos los métodos no paramétricos, no necesitamos hacer ninguna hipótesis acerca de la distribución que siguen nuestros datos. Se trata de un estadígrafo que se calcula de un modo similar al coeficiente de Pearson, pero asignando rangos a los datos. Cuando hay varios datos con el mismo valor se asigna a todos ellos el promedio de los rangos correspondientes.

Este coeficiente funciona para cualquier tipo de asociación entre variables que siga una función de tipo monotónico. Obviamente, esto se debe al hecho de utilizar una escala ordinal, con lo que esta propiedad sirve para cualquier otro coeficiente que emplee rangos para ser calculado. Esta es una diferencia importante frente al coeficiente de Pearson, que sólo responde bien si la relación entre ambas variables es lineal. Además, sirve tanto para variables cuantitativas como para cualitativas ordinales.

Se calcula por medio de la siguiente fórmula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

donde d_i es la diferencia, para cada dato u observación, entre los rangos que toman los valores de las dos variables.

El rango de valores que toma este coeficiente va desde -1 a 1. Cuanto más cerca de estos extremos, más fuerte es el grado de asociación, y el signo indica si la relación funcional entre ambas variables es descendente o ascendente, de la misma forma que el coeficiente de Pearson.

Para determinar la significación de este coeficiente se utiliza el siguiente estadístico, cuya distribución se aproxima a la *t* de Student cuando el tamaño de muestra es mayor que 20 unidades:

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

Al igual que con el coeficiente de Pearson, se establece como H_0 que $\rho = 0$. En el Cuadro X.1 se muestra un ejemplo del cálculo de este coeficiente.

X.1.3. COEFICIENTE τ DE KENDALL

Este coeficiente es especialmente indicado para variables expresadas de forma ordinal. Por lo tanto también se puede utilizar tras asignar rangos a los valores de las variables.

Existe una fórmula para su cálculo que no tiene en cuenta "empates" entre rangos de datos, y que se denota por τ_A, y que se calcula de la siguiente manera:

$$\tau_A = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

donde n_c es el número de pares de rangos concordantes y n_d es el número de pares de rangos discordantes.

En el caso de que tengamos datos en los cuales existen "empates" entre rangos, utilizaríamos esta otra fórmula:

$$\tau_B = \frac{n_p - n_q}{\sqrt{(n_p + n_q + n_{E(X)}) \cdot (n_p + n_q + n_{E(Y)})}}$$

donde *p* indica concordancia o no inversión, *q* discordancia o inversión y *E* empate, considerando todos los pares de casos.

Los valores de este coeficiente se encuentran entre -1 y 1, y sus propiedades e interpretación son análogas a los coeficientes de Pearson y Spearman.

La significación se contrasta por medio de estadísticos que se distribuyen siguiendo una normal. En el caso de τ_A tendríamos la siguiente fórmula:

$$z = \frac{3(n_c - n_d)}{\sqrt{\frac{n(n-1)(2n+5)}{2}}}$$

Para el caso de τ_B tendríamos la siguiente fórmula:

$$z = \frac{n_c - n_d}{\sqrt{\upsilon}}$$

donde:

$$\upsilon = \frac{\upsilon_0 - \upsilon_t - \upsilon_u}{18 + \upsilon_1 + \upsilon_2}$$

donde a su vez:

$$\upsilon_0 = n(n-1)(2n+5)$$
$$\upsilon_t = \sum t_i(t_i - 1)(2t_i + 5)$$
$$\upsilon_u = \sum_j u_j(u_j - 1)(2u_j + 5)$$
$$\upsilon_1 = \sum_i t_i(t_i - 1)\sum_j \frac{u_j(u_j - 1)}{2n(n-1)}$$
$$\upsilon_2 = \sum_i t_i(t_i - 1)(t_i - 2)\sum_j \frac{u_j(u_j - 1)(u_j - 2)}{9n(n-1)(n-2)}$$

La H_0 para la significación de este coeficiente es que su valor es 0 y, por tanto, no hay asociación entre variables. En el Cuadro X.1 se muestran ejemplos del cálculo de este coeficiente.

X.1.4. COEFICIENTE GAMMA (γ)

Es el coeficiente más adecuado cuando los datos presentan un alto solapamiento de rangos. De nuevo, el intervalo de valores que puede tomar este coeficiente es entre -1 y 1, siendo -1 asociación perfecta negativa, 1 asociación perfecta positiva y 0 no asociación.

La fórmula para su cálculo es la siguiente:

$$\gamma = \frac{n_c - n_d}{n_c + n_d}$$

donde n_c es el número de pares de rangos concordantes y n_d es el número de pares discordantes.

La significación se comprueba utilizando el valor del siguiente estadístico, que se distribuye siguiendo una Normal:

$$z = \gamma\sqrt{\frac{n_c - n_d}{N(1-\gamma^2)}}$$

Al igual que con el resto de coeficientes, la H_0 que se establece es que el valor del coeficiente es igual a 0, y por lo tanto no existe asociación entre las variables.

En el Cuadro X.1 se muestra un ejemplo del cálculo de este coeficiente.

CUADRO X.1. Correlaciones bivariadas simples

Aplicación con R.

EJEMPLO. En el **Cuadro X.1.V.csv** tenemos datos de un muestreo realizado en nidadas de una pequeña especie de ave, la codorniz común. En este muestreo se contó el número de huevos por hembra, puestos durante una época de cría, en 50 hembras marcadas cuya edad era conocida. Al mismo tiempo se registraron la temperatura media durante la misma época de cría y la disponibilidad de alimento para las hembras (cuyos valores son porcentajes respecto al máximo medido). El mismo muestreo se repitió en la siguiente época de cría con una muestra diferente de hembras en la misma población. El objetivo de la investigación es descubrir si existe relación entre la edad de la hembra y el número de huevos por puesta.

Después de invocar al programa R y seleccionar el directorio de trabajo como se explica en el Apéndice II, debemos abrir el *script* **Cuadro X.1.R** que se muestra en la siguiente ventana. Para este *script* es necesario tener instalados, siguiendo las instrucciones del Apéndice II, los paquetes "ltm", "MASS", "msm", "mvtnorm", "polycor" y "sfsmisc".

```
#Cuadro X.1
#Correlaciones bivariadas simples.

#########################################################
#Sección modificable por el usuario
#########################################################

#Lectura de la base de datos
datos<-read.csv2("Cuadro X.1.V.csv",header=T,encoding="latin1")

#Selección de variables de interés.
varInteresX<-c("Edad.hembra")
varInteresY<-c("N°.huevos")

#Selección de variables de agrupación.
#varAgrupacion<-NULL
varAgrupacion<-c("Provincia")
#varAgrupacion<-c("Localidad")
#varAgrupacion<-c("Provincia","Localidad")

#Pruebas a realizar
#1.Coeficiente de correlación de Pearson
#2.Coeficiente de correlación de Spearman
#3.Coeficiente de correlación de Kendall

pruebas<-c(1,2,3)

#Nombre del archivo de salida
ArchivodeSalida<-"Salida Cuadro X.1.txt"
```

CUADRO X.1. *(Continuación)*[R]

Paso 1. En primer lugar es necesario importar los datos desde nuestro archivo csv.

```
datos<-read.csv2("Cuadro X.1.V.csv",header=T,encoding="latin1")
```

Paso 2. Seleccionamos las variables de interés para nuestros análisis. Las variables *X* e *Y* serán independiente y dependiente, respectivamente, aunque son intercambiables, el valor del coeficiente no depende de que la variable dependiente sea la primera o la segunda.

```
varInteresX<-c("Edad.hembra")
varInteresY<-c("Nº.huevos")
```

Paso 3. Posteriormente, en función de si nos interesa calcular un coeficiente de correlación para cada grupo de datos que tengamos (por Provincia, por Localidad, o ambos) seleccionamos si queremos agrupación o no, y de qué tipo. En este caso, seleccionamos agrupar en función de la «Provincia».

```
#varAgrupacion<-NULL
varAgrupacion<-c("Provincia")
#varAgrupacion<-c("Localidad")
#varAgrupacion<-c("Provincia","Localidad")
```

Paso 4. Entre todas las pruebas que podemos realizar con este script, seleccionamos las 3 pruebas. El coeficiente de Pearson es solo para variables cuantitativas, pero los otros dos se pueden usar con variables cuantitativas o cualitativas. Los tres coeficientes cuantifican el grado de correlación entre las variables.

```
#1.Coeficiente de correlación de Pearson
#2.Coeficiente de correlación de Spearman
#3.Coeficiente de correlación de Kendall

pruebas<-c(1,2,3)
```

Paso 5. Por último, se especifica el nombre del archivo de salida.

```
ArchivodeSalida<-"Salida Cuadro X.1.txt"
```

Paso 6. Ahora seleccionamos el contenido de este *script* pulsando *ctrl+A* y lo ejecutamos pulsando *ctrl+R*. A continuación abrimos el archivo de salida, para ver los resultados.

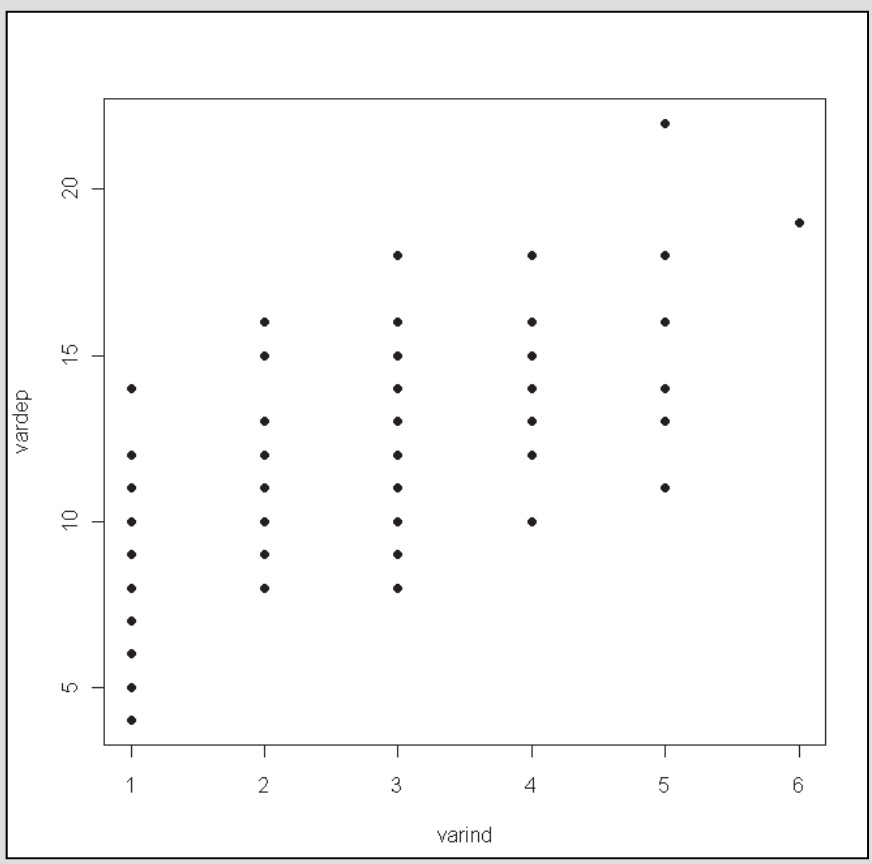
CUADRO X.1. *(Continuación)*[R]

```
[[1]]
[1] "Coeficiente de Pearson"

$Palencia

       varind vardep
varind  *****  0.448
vardep  0.004  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values


$Soria

       varind vardep
varind  *****  0.777
vardep <0.001  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values


[[4]]
[1] "Coeficiente de Spearman"

$Palencia

       varind vardep
varind  *****  0.352
vardep  0.026  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values


$Soria

       varind vardep
varind  *****  0.775
vardep <0.001  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values


[[7]]
[1] "Coeficiente de Kendall"

$Palencia

       varind vardep
varind  *****  0.282
vardep  0.022  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values


$Soria

       varind vardep
varind  *****  0.663
vardep <0.001  *****

upper diagonal part contains correlation coefficient estimates
lower diagonal part contains corresponding p-values
```

CUADRO X.1. *(Continuación)*[R]

Se observa que, en todos los casos, se rechaza la hipótesis de independencia, ya que la probabilidad es siempre menor que 0,05. Por lo tanto, en ambas provincias y considerando los tres coeficientes, se concluye que existe una relación de dependencia entre la edad de la hembra y el número de huevos. Esta asociación parece ser más fuerte en la provincia de Soria donde, por ejemplo, el coeficiente de correlación de Pearson es 0,78, mientras que en Palencia es 0,45. Los coeficientes de Spearman y Kendall también muestran esta mayor asociación entre ambas variables en una provincia que en la otra.

Paso 7. En los resultados también aparece una gráfica representando todos los datos de las variables *X* e *Y*. Esta representación permite ver si la relación es lineal. Es importante mencionar que se representan todos los datos de ambas variables, es decir, aunque se seleccionara realizar agrupación –como fue en este caso– no se muestran los distintos grupos en la gráfica.

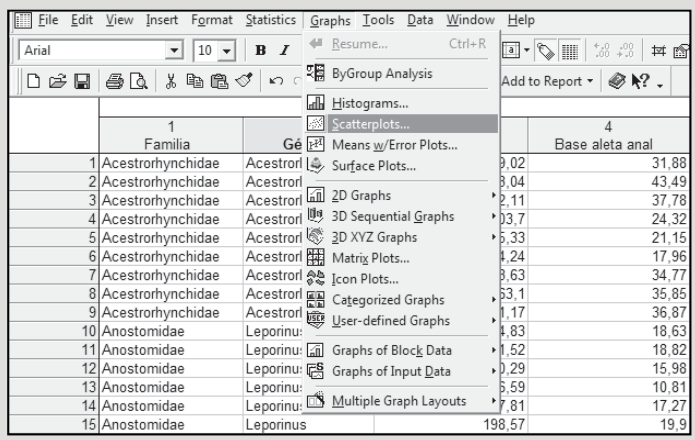


CUADRO X.1. *(Continuación)*[STATISTICA]

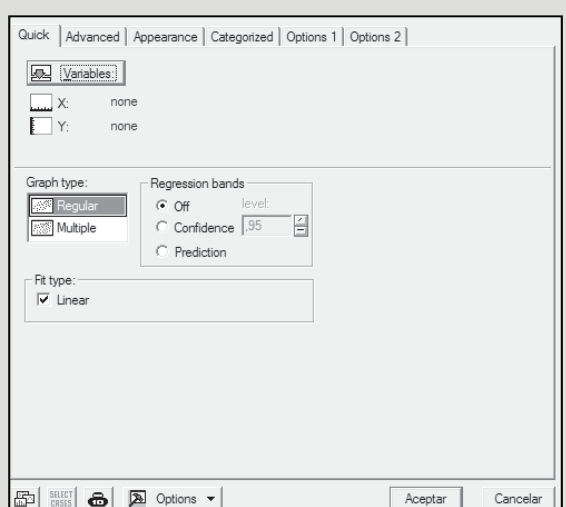
Aplicación con STATISTICA.

EJEMPLO. En el **Cuadro X.1.sta** hay datos de la longitud estándar y la longitud de la base de la aleta anal de diferentes especies de peces. El objetivo es determinar si existe una relación entre la longitud estándar y la longitud de la aleta anal. En primer lugar calcularemos el coeficiente de Pearson, después Spearman, Kendall y Gamma. El coeficiente de Pearson es solo para variables cuantitativas, pero los otros dos se pueden usar con variables cuantitativas o cualitativas. Los tres coeficientes cuantifican el grado de correlación entre las variables.

Paso 1. En primer lugar representamos los datos para ver el tipo de relación entre las variables. Para ello seleccionamos «Graphs» ⇨ «Scatterplots».

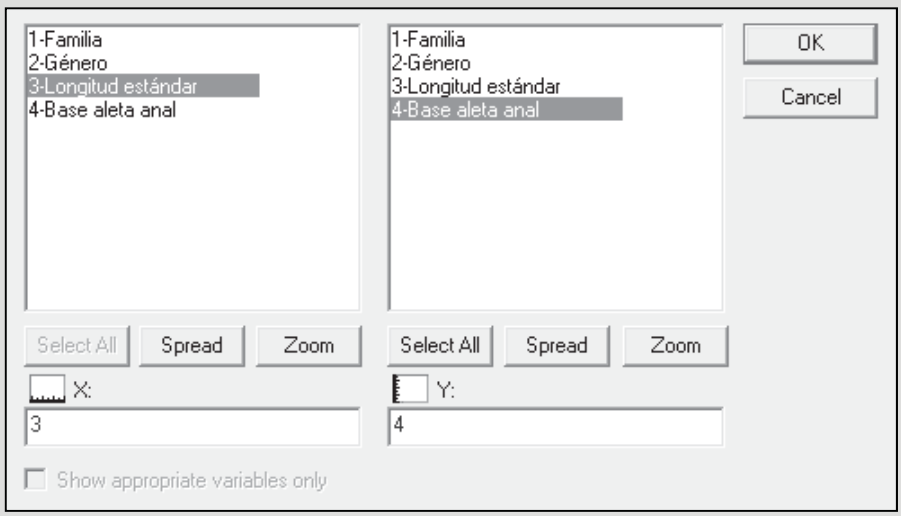


Paso 2. En la siguiente ventana escogemos la opción «Variables».

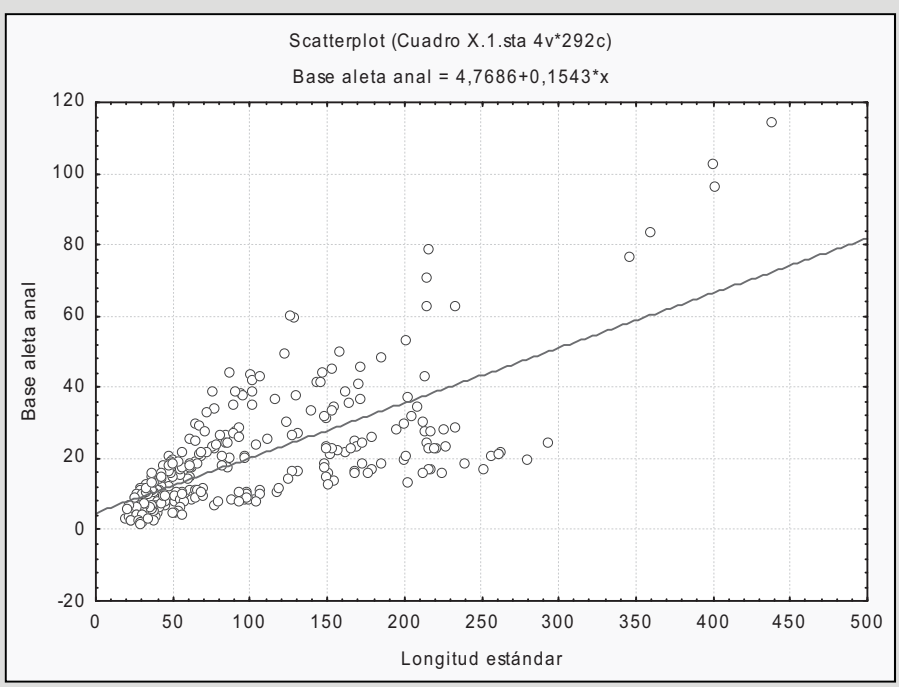


CUADRO X.1. *(Continuación)* *STATISTICA*

Paso 3. Se seleccionan las dos variables a representar.

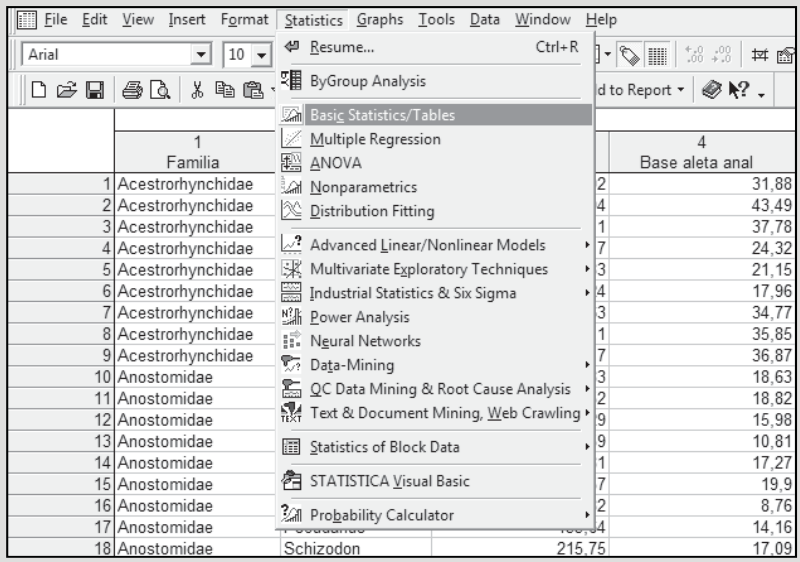


Paso 4. Se observa que parece existir una relación lineal entre ambas variables.

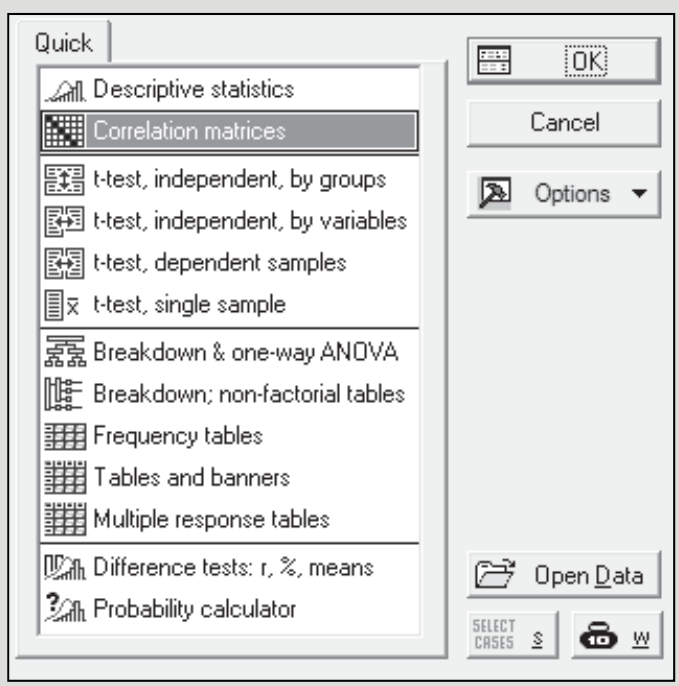


CUADRO X.1. *(Continuación)*[STATISTICA]

Paso 5. A continuación se selecciona en el menú principal «Statistics» ⇨ «Basic Statistics/Tables».

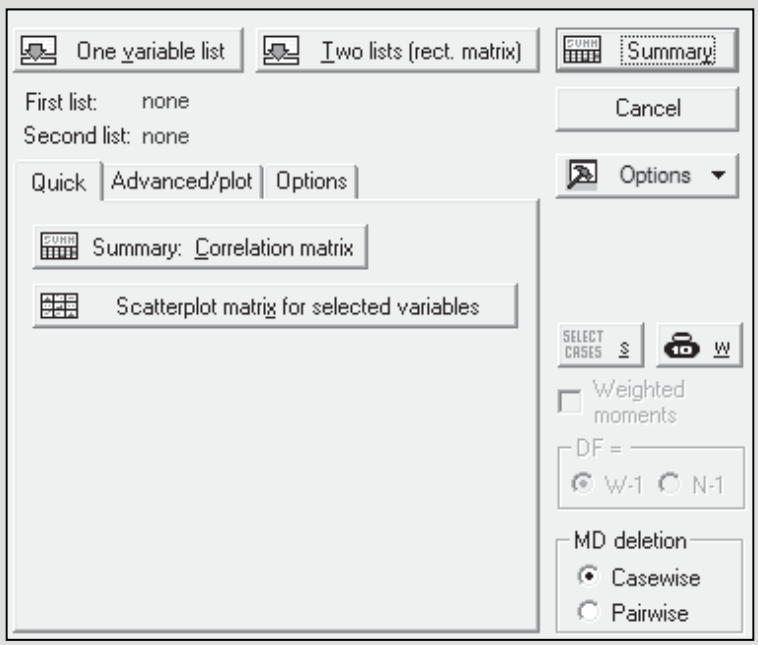


Paso 6. En la siguiente ventana escogemos la opción «Correlation matrices» y pulsamos «OK».

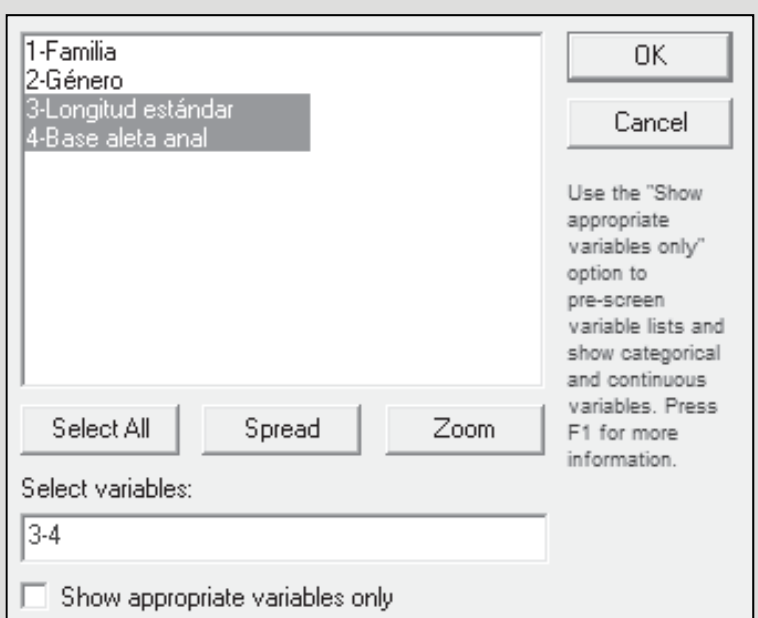


CUADRO X.1. *(Continuación)*[STATISTICA]

Paso 7. Aparece la siguiente ventana, donde escogemos las variables en la pestaña «One variable list».

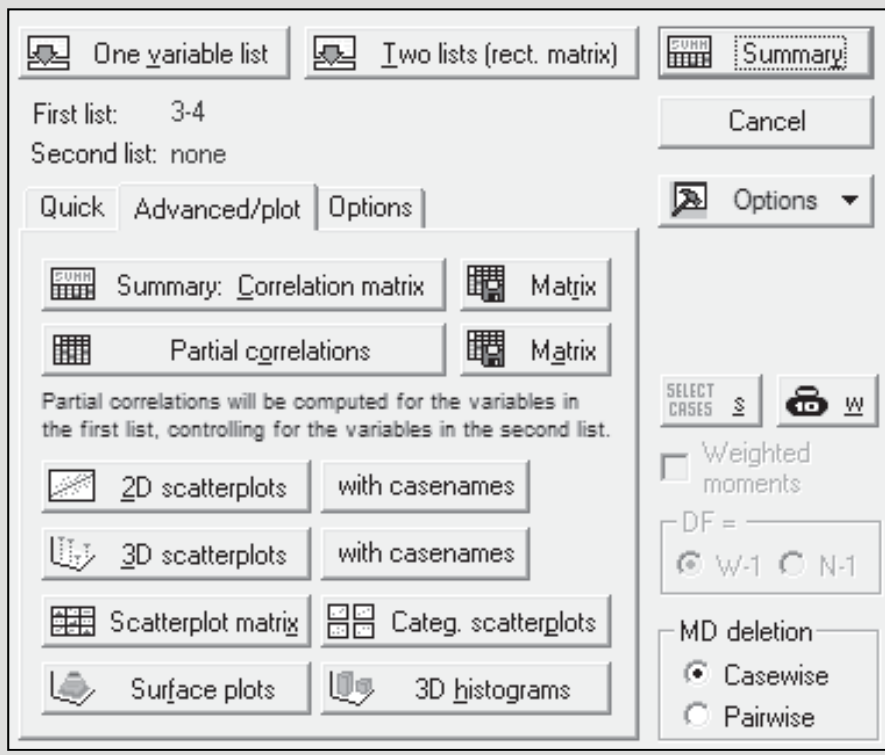


Paso 8. Seleccionamos «Longitud estándar» y «Base aleta anal».



CUADRO X.1. *(Continuación)*[STATISTICA]

Paso 9. En la ventana que aparece nos situamos en la pestaña «Advanced/plot» y pulsamos en «Summary. Correlation matrix» para obtener los resultados.

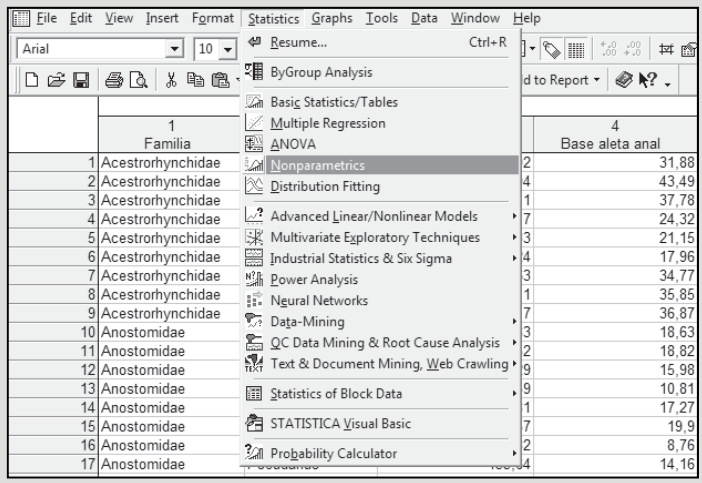


Paso 10. El coeficiente de correlación de Pearson (0,71) muestra una correlación positiva y significativa ($p < 0,05$) entre la longitud estándar de los peces y la longitud de la base de la aleta anal.

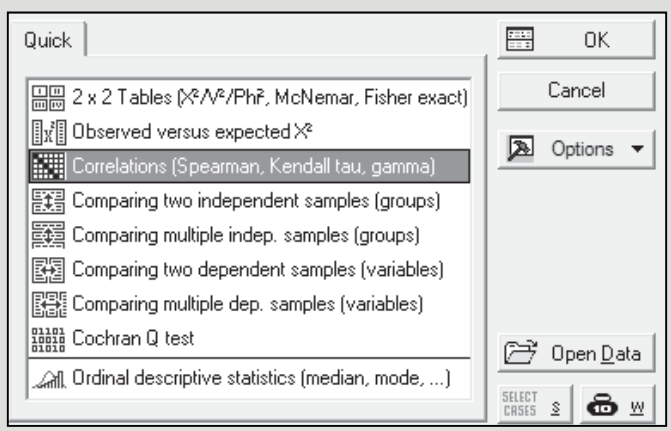
Variable	Correlations (Cuadro X.1.sta) Marked correlations are significant at p < ,05000 N=292 (Casewise deletion of missing data) Longitud estándar	Base aleta anal
Longitud estándar	1,00	0,71
Base aleta anal	0,71	1,00

Paso 11. Ahora, para calcular los coeficientes no paramétricos mencionados al comienzo, debemos proceder con los siguientes pasos. En primer lugar vamos al menú «Statistics» ⇨ «Nonparametrics».

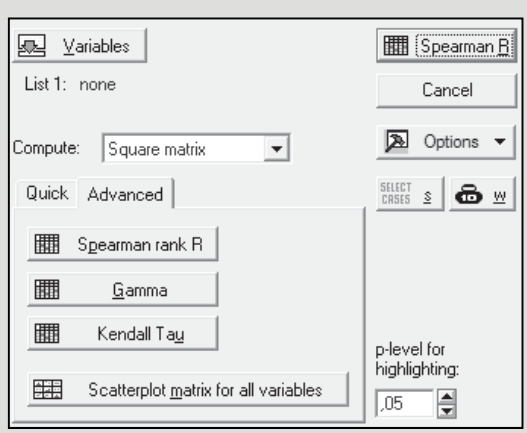
CUADRO X.1. *(Continuación)* *STATISTICA*



Paso 12. En el siguiente cuadro, seleccionamos la opción «Correlations».

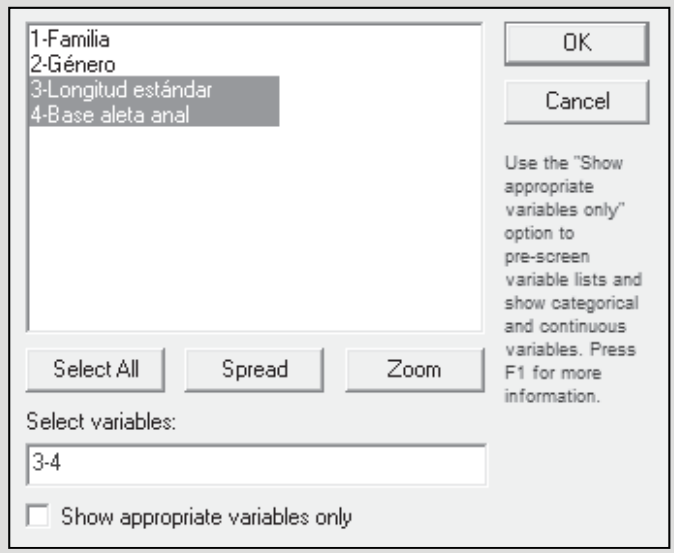


Paso 13. Pulsamos «OK» y en el siguiente cuadro nos vamos a la pestaña «Advanced».



CUADRO X.1. *(Continuación)*^{STATISTICA}

Paso 14. Ahora debemos seleccionar las variables, en el botón «Variables». Seleccionaremos las variables «Longitud estándar» y «Base aleta anal».



Paso 15. Pulsamos «OK» y volviendo al cuadro del paso 3, podemos ver los resultados de cada una de las pruebas si pulsamos en «Spearman rank R», «Gamma» y «Kendall Tau».

Variable	Spearman Rank Order Correlations (Cuadro X.1.sta) MD pairwise deleted Marked correlations are significant at p <,05000	
	Longitud estándar	Base aleta anal
Longitud estándar	1,000000	0,732605
Base aleta anal	0,732605	1,000000

Variable	Gamma Correlations (Cuadro X.1.sta) MD pairwise deleted Marked correlations are significant at p <,05000	
	Longitud estándar	Base aleta anal
Longitud estándar	1,000000	0,530481
Base aleta anal	0,530481	1,000000

Variable	Kendall Tau Correlations (Cuadro X.1.sta) MD pairwise deleted Marked correlations are significant at p <,05000	
	Longitud estándar	Base aleta anal
Longitud estándar	1,000000	0,530375
Base aleta anal	0,530375	1,000000

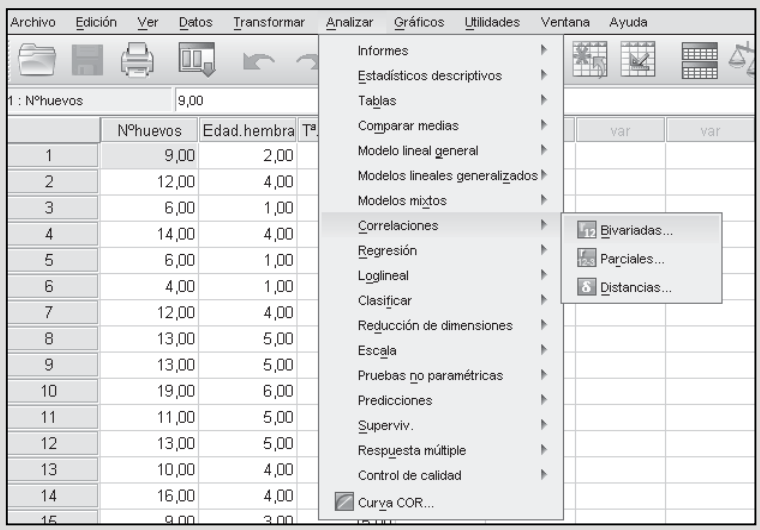
Todas las correlaciones son positivas y significativas ($p < 0,05$). Por tanto, se concluye que existe una relación significativa entre la longitud estándar y la longitud de la base de la aleta anal, aunque los valores del coeficiente Gamma y la Tau de Kendall no son muy altos.

CUADRO X.1. *(Continuación)*[SPSS]

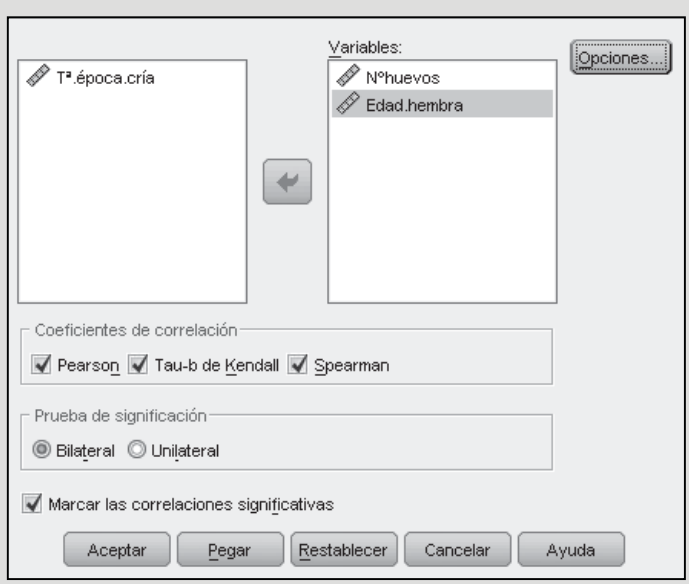
Aplicación con SPSS.

EJEMPLO. Volveremos a utilizar los mismos datos del ejemplo con R, que se encuentran también en formato SPSS en el archivo **Cuadro X.1.s**av.

Paso 1. Abrimos el archivo de datos **Cuadro X.1.sav** y vamos al menú «Analizar» ⇨ «Correlaciones» ⇨ «Bivariadas».



Paso 2. En el cuadro de diálogo que aparece seleccionamos las variables «Nºhuevos» y «Edad.hembra» y con ayuda de la flecha, la pasamos al recuadro de la derecha. Seleccionamos, además de «Pearson», «Tau-b de Kendall» y «Spearman».



CUADRO X.1. *(Continuación)*[SPSS]

Paso 3. Pulsamos en «Aceptar» y obtendremos el siguiente cuadro con los resultados. Nos muestra por separado las correlaciones paramétricas y las no paramétricas.

Correlaciones

		Nºhuevos	Edad.hembra
Nºhuevos	Correlación de Pearson	1	,637**
	Sig. (bilateral)		,000
	N	100	100
Edad.hembra	Correlación de Pearson	,637**	1
	Sig. (bilateral)	,000	
	N	100	100

**. La correlación es significativa al nivel 0,01 (bilateral).

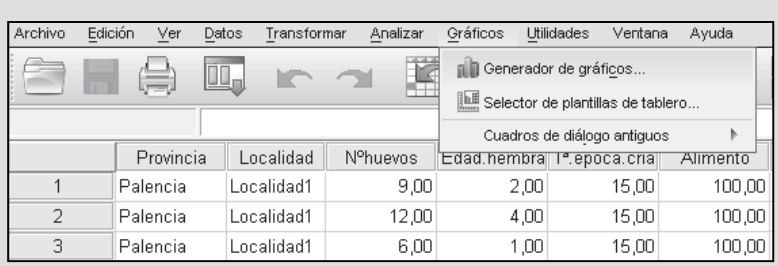
Correlaciones no paramétricas

			Nºhuevos	Edad.hembra
Tau_b de Kendall	Nºhuevos	Coeficiente de correlación	1,000	,498**
		Sig. (bilateral)	.	,000
		N	100	100
	Edad.hembra	Coeficiente de correlación	,498**	1,000
		Sig. (bilateral)	,000	.
		N	100	100
Rho de Spearman	Nºhuevos	Coeficiente de correlación	1,000	,609**
		Sig. (bilateral)	.	,000
		N	100	100
	Edad.hembra	Coeficiente de correlación	,609**	1,000
		Sig. (bilateral)	,000	.
		N	100	100

**. La correlación es significativa al nivel 0,01 (bilateral).

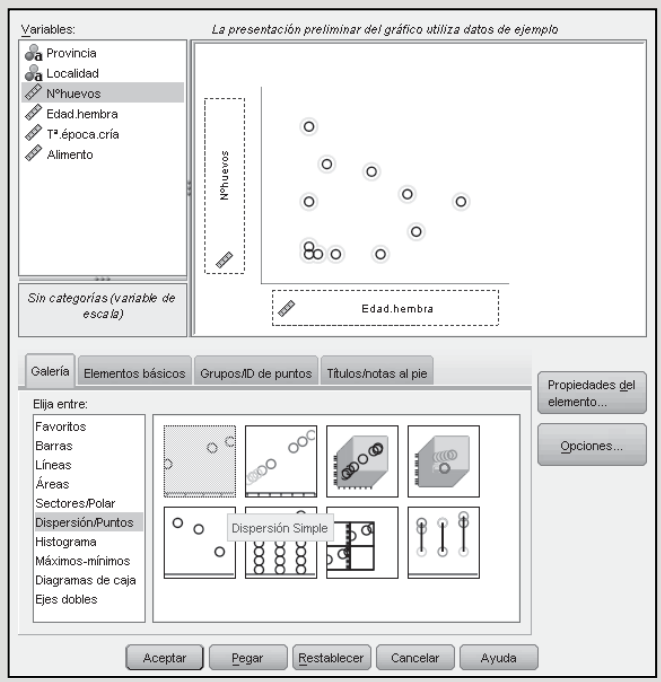
Los tres coeficientes muestran que existe una correlación significativa y positiva entre ambas variables (en todos los casos la $p < 0{,}01$). El número de huevos es mayor a medida que aumenta la edad de la hembra. El coeficiente de Pearson muestra que el grado de asociación entre ambas variables es de 0,64. Es importante recordar que el coeficiente de Pearson es solo para variables cuantitativas, pero los otros dos se pueden usar con variables cuantitativas o cualitativas.

Paso 4. Debido a que los tres coeficientes cuantifican el grado de correlación entre las variables, es conveniente representar los datos para ver el tipo de asociación entre la edad de la hembra y el número de huevos. Para ello vamos a «Gráficos» ⇨ «Generador de gráficos».

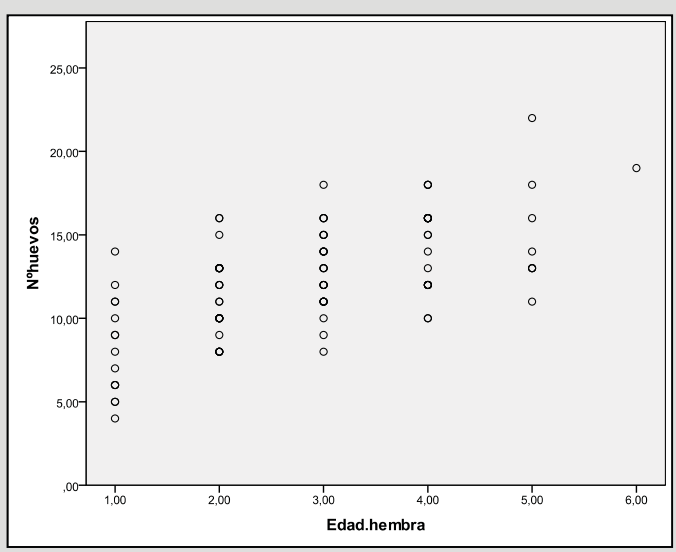


CUADRO X.1. *(Continuación)*[SPSS]

Paso 5. En «Galería» se selecciona «Dispersión/Puntos» y entre los distintos gráficos que aparecen se escoge «Dispersión Simple». Posteriormente se arrastra desde el cuadro «Variables» a los ejes *X* e *Y* de la gráfica la variable correspondiente.



La representación de los datos muestra que existe una clara relación lineal entre ambas variables.



X.2. Correlaciones bivariadas parciales

X.2.1. COEFICIENTE DE CORRELACIÓN PARCIAL DE PEARSON

En ocasiones, una correlación lineal bivariada puede estar enmascarada por el efecto de otras variables que no controlamos. Por ejemplo, puede existir una relación fuerte entre el tamaño de la nidada de una determinada especie de pájaro y la edad de la hembra y, sin embargo, esta correlación se enmascara por la influencia en nuestros datos de otras variables como podrían ser la temperatura media durante la época de cría. Para estudiar una correlación simple entre dos variables de interés, a la vez que controlamos el efecto de otras variables, utilizamos las correlaciones parciales.

Si tenemos dos variables *X* e *Y* sobre las que queremos calcular la correlación parcial controlando *k* variables *Z*, el método se basa en calcular el coeficiente de correlación de Pearson entre dos variables derivadas d_x y d_y tales que:

$$d_x = \textit{residuos de la regresión lineal de x con } z_1, \ldots, z_k$$
$$d_y = \textit{residuos de la regresión lineal de y con } z_1, \ldots, z_k$$

Es posible controlar al mismo tiempo el efecto de una o más variables *Z*. Si controlamos sólo una, hablaríamos de correlación parcial de primer orden. Si controlamos dos, correlación parcial de segundo orden, y así sucesivamente.

El cálculo de las correlaciones parciales de primer orden se realiza por medio de la siguiente fórmula:

$$r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

donde *r* es el coeficiente de correlación de Pearson entre las variables indicadas.

Como ejemplo de mayor orden, mostramos la fórmula del coeficiente en el caso de que se estén controlando dos variables (z_i y z_j) con lo cual es una correlación parcial de segundo orden. Las de mayor orden se obtendrían siguiendo la misma lógica:

$$r_{xy,z_iz_j} = \frac{r_{xy,z_i} - r_{xz_j,z_i}\, r_{yz_j,z_i}}{\sqrt{(1 - r_{xz_j,z_i}^2)(1 - r_{yz_j,z_i}^2)}}$$

La prueba de significación de este coeficiente, con la hipótesis nula de que su valor poblacional es cero, se realiza calculando el estadístico *t* por medio de la siguiente fórmula:

$$T = r_{xy,k}\frac{\sqrt{m - k - 2}}{\sqrt{(1 - r_{xy,k}^2)}}$$

donde *m* es el número de casos (si no es igual para todos los pares de variables debe tomarse el menor de ellos) y *k* es el número de variables controladas que hemos empleado.

En el Cuadro X.2 se muestra un ejemplo de cálculo de correlaciones parciales.

X.2.2. COEFICIENTE DE CORRELACIÓN PARCIAL DE KENDALL ($T_{xy.z}$)

El fundamento de las correlaciones parciales es aquí el mismo que en el caso de la estadística paramétrica (ver apartado X.2.1) pero necesitamos un estadístico diferente, el cual nos permita trabajar sin hacer ninguna asunción acerca de la distribución que siguen nuestros datos.

El principal estadístico que cumple esta función es la *T* de Kendall, que es en esencia la misma que la τ descrita en el apartado X.1.3 adaptada al cálculo de correlaciones parciales.

La fórmula para computarla es la siguiente:

$$T_{xy.z} = \frac{T_{xy} - T_{xz}T_{yz}}{\sqrt{(1 - T_{xz}^2)(1 - T_{yz}^2)}}$$

donde cada correlación bivariada *T* se calcula de la misma forma que la τ de Kendall del apartado X.1.3.

Para el cálculo de su significación se emplea el siguiente estadístico, que sigue una distribución Normal:

$$z = \frac{3T_{xy.z}\sqrt{N(N-1)}}{\sqrt{2(2N+5)}}$$

La H_0 que se establece, como es habitual para estos estadísticos, es que el valor del coeficiente poblacional es 0.

En el Cuadro X.2 se muestra un ejemplo del cálculo de este coeficiente.

CUADRO X.2. Correlaciones bivariadas parciales

Aplicación con R.

EJEMPLO. En el ejemplo del Cuadro X.1 en la aplicación con R se observó que había una correlación positiva entre el número de huevos por hembra y la edad de las hembras. Sin embargo, está relación puede variar en función del alimento y/o la temperatura durante la época de cría. Para determinar si efectivamente está relación está afectada por otras variables es necesario estudiar la correlación parcial.

Después de invocar al programa R y seleccionar el directorio de trabajo como se explica en el Apéndice II, debemos abrir el *script* **Cuadro X.2.R** que se muestra en la siguiente ventana. Para este script necesitaremos instalar, siguiendo las instrucciones del Apéndice II, los paquetes "tcltk", "ggm", "car" y "Rcmdr".

CUADRO X.2. *(Continuación)*[R]

```
#Cuadro X.2
#Correlaciones bivariadas parciales.

#########################################################
#Sección modificable por el usuario
#########################################################

#Lectura de la base de datos
datos<-read.csv2("Cuadro X.1.V.csv",header=T,encoding="latin1")

#Selección de variables de interés.
varInteresX<-c("Edad.hembra")
varInteresY<-c("N°.huevos")
#varInteresZ<-c("Tª.epoca.cria")
#varInteresZ<-c("Alimento")
#varInteresZ<-c("Tª.epoca.cria","Alimento")
varInteresZ<-NULL

#Nombre del archivo de salida
sink("Salida Cuadro X.2.txt")
```

Paso 1. En primer lugar es necesario importar los datos desde nuestro archivo csv. Será el mismo fichero que en el ejemplo del Cuadro X.1.

datos<-read.csv2("Cuadro X.1.V.csv",header=T,encoding="latin1")

Paso 2. Seleccionamos las variables de interés para nuestros análisis. Las variables *X* e *Y* serán independiente y dependiente, respectivamente. La variable *Z* es aquella que queremos controlar para calcular el coeficiente de correlación parcial. En primer lugar, no utilizaremos ninguna variable control. Por ello especificamos como «NULL» la «varInteresZ».

varInteresX<-c("Edad.hembra")
varInteresY<-c("Nº.huevos")
#varInteresZ<-c("Tª.epoca.cria")
#varInteresZ<-c("Alimento")
#varInteresZ<-c("Tª.epoca.cria","Alimento")
varInteresZ<-NULL

Paso 3. No es necesario seleccionar las pruebas porque se calcula tanto el coeficiente de correlación parcial de Pearson como el coeficiente de correlación parcial de Kendall.

Paso 4. Se especifica el nombre del archivo de salida.

ArchivodeSalida<-"Salida Cuadro X.2.txt"

Paso 5. Ahora seleccionamos el contenido de este *script* pulsando *Ctrl+A* y lo ejecutamos pulsando *Ctrl+R*. A continuación abrimos el archivo de salida, para ver los resultados.

CUADRO X.2. *(Continuación)*[R]

```
[1] "Coeficiente de Pearson"
          varind    vardep
varind 0.0000000 0.6371661
vardep 0.6371661 0.0000000
$tval
         varind   vardep
varind 0.000000 8.183997
vardep 8.183997 0.000000

$df
[1] 98

$pvalue
              varind       vardep
varind 1.000000e+00 1.025775e-12
vardep 1.025775e-12 1.000000e+00

[1] "Coeficiente de Kendall"
          varind    vardep
varind 0.0000000 0.4983081
vardep 0.4983081 0.0000000
$tval
         varind   vardep
varind 0.000000 5.689733
vardep 5.689733 0.000000

$df
[1] 98

$pvalue
              varind       vardep
varind 1.000000e+00 1.322216e-07
vardep 1.322216e-07 1.000000e+00
```

En el coeficiente de Pearson, la primera matriz muestra que existe una correlación entre ambas variables de 0,637, la segunda matriz da el valor del estadístico, los grados de libertad son 98 y la última matriz es la probabilidad. Se observa que se rechaza la hipótesis de independencia, ya que $p < 0{,}001$. Por tanto, existe una correlación significativa entre la edad de la hembra y el número de huevos por hembra. En el coeficiente de Kendall es 0,498 y también $p < 0{,}001$.

Paso 6. Repetimos el proceso pero está vez incluyendo la variable «Tª.epoca.cria», como variable control.

```
varInteresX<-c("Edad.hembra")
varInteresY<-c("Nº.huevos")
varInteresZ<-c("Tª.epoca.cria")
#varInteresZ<-c("Alimento")
#varInteresZ<-c("Tª.epoca.cria","Alimento")
#varInteresZ<-NULL
```

CUADRO X.2. *(Continuación)*[R]

Paso 7. De nuevo solo hay que fijarse en el coeficiente entre la variable dependiente y la independiente, que para Pearson es 0,842 y para Kendall 0,622. En ambos casos la probabilidad, que se ve en la matriz «pvalue», es significativa para las dos pruebas con $p < 0{,}001$. Los grados de libertad son 97. Por tanto, como ambos coeficientes aumentan con respecto a cuando no se consideraba la variable control (véase paso 5), se concluye que la relación entre la edad de la hembra y el número de huevos por hembra varía en función de la temperatura durante la época de cría.

```
[1] "Coeficiente de Pearson"
                varind    vardep Tª.epoca.cria
varind         0.0000000 0.8422472    -0.7185677
vardep         0.8422472 0.0000000     0.8133817
Tª.epoca.cria -0.7185677 0.8133817     0.0000000
$tval
                varind   vardep Tª.epoca.cria
varind          0.00000 15.38732     -10.17615
vardep         15.38732  0.00000      13.77078
Tª.epoca.cria -10.17615 13.77078       0.00000

$df
[1] 97

$pvalue
                    varind       vardep Tª.epoca.cria
varind        1.000000e+00 8.942973e-28  5.586822e-17
vardep        8.942973e-28 1.000000e+00  1.489113e-24
Tª.epoca.cria 5.586822e-17 1.489113e-24  1.000000e+00

[1] "Coeficiente de Kendall"
                varind    vardep Tª.epoca.cria
varind         0.0000000 0.6220117     -0.435895
vardep         0.6220117 0.0000000      0.618560
Tª.epoca.cria -0.4358950 0.6185601      0.000000
$tval
                varind   vardep Tª.epoca.cria
varind         0.000000 7.823809     -4.770089
vardep         7.823809 0.000000      7.753365
Tª.epoca.cria -4.770089 7.753365      0.000000

$df
[1] 97

$pvalue
                    varind       vardep Tª.epoca.cria
varind        1.000000e+00 6.342434e-12  6.492482e-06
vardep        6.342434e-12 1.000000e+00  8.936207e-12
Tª.epoca.cria 6.492482e-06 8.936207e-12  1.000000e+00
```

Paso 8. Repetimos el proceso pero está vez incluyendo las dos variables control «Tª.epoca.cria» y «Alimento».

CUADRO X.2. *(Continuación)*[R]

```
varInteresX<-c("Edad.hembra")
varInteresY<-c("Nº.huevos")
#varInteresZ<-c("Tª.epoca.cria")
#varInteresZ<-c("Alimento")
varInteresZ<-c("Tª.epoca.cria","Alimento")
#varInteresZ<-NULL
```

Paso 9. Hay que fijarse solo en el coeficiente entre la variable dependiente y la independiente, que para Pearson es 0,839 y para Kendall 0,62. En ambos casos la probabilidad, que se ve en la matriz «pvalue», es significativa para las dos pruebas con $p < 0{,}001$. Los grados de libertad son 96. Sin embargo, los coeficientes no son mayores con respecto a cuando se consideraba la variable control (temperatura durante la época de cría, véase paso 7). Por tanto, se concluye que la relación entre la edad de la hembra y el número de huevos por hembra no varía en función del alimento.

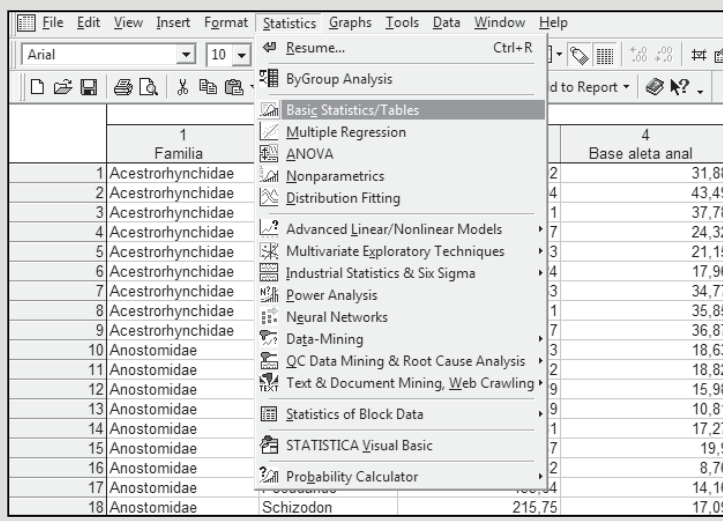
```
[1] "Coeficiente de Pearson"
                  varind     vardep Tª.epoca.cria   Alimento
varind         0.00000000 0.83925099    -0.7053061 0.02893592
vardep         0.83925099 0.00000000     0.7779235 0.05199749
Tª.epoca.cria -0.70530613 0.77792354     0.0000000 0.23253662
Alimento       0.02893592 0.05199749     0.2325366 0.00000000
$tval
                  varind     vardep Tª.epoca.cria  Alimento
varind         0.0000000 15.1228216     -9.748247 0.2836317
vardep        15.1228216  0.0000000     12.130110 0.5101594
Tª.epoca.cria -9.7482474 12.1301098      0.000000 2.3426004
Alimento       0.2836317  0.5101594      2.342600 0.0000000

$df
[1] 96

$pvalue
                  varind       vardep Tª.epoca.cria   Alimento
varind        1.000000e+00 3.817921e-27  5.167167e-16 0.77730342
vardep        3.817921e-27 1.000000e+00  4.403922e-21 0.61111075
Tª.epoca.cria 5.167167e-16 4.403922e-21  1.000000e+00 0.02121476
Alimento      7.773034e-01 6.111107e-01  2.121476e-02 1.00000000

[1] "Coeficiente de Kendall"
                  varind    vardep Tª.epoca.cria   Alimento
varind         0.0000000 0.6207580    -0.4004457 -0.0230591
vardep         0.6207580 0.0000000     0.5393389  0.1066714
Tª.epoca.cria -0.4004457 0.5393389     0.0000000  0.3483801
Alimento      -0.0230591 0.1066714     0.3483801  0.0000000
$tval
                  varind   vardep Tª.epoca.cria   Alimento
varind         0.0000000 7.757837     -4.281854 -0.2259922
vardep         7.7578370 0.000000      6.275373  1.0511596
Tª.epoca.cria -4.2818537 6.275373      0.000000  3.6415445
Alimento      -0.2259922 1.051160      3.641545  0.0000000

$df
[1] 96

$pvalue
                  varind       vardep Tª.epoca.cria    Alimento
varind        1.000000e+00 9.213063e-12  4.394710e-05 0.821687943
vardep        9.213063e-12 1.000000e+00  1.004510e-08 0.295824123
Tª.epoca.cria 4.394710e-05 1.004510e-08  1.000000e+00 0.000439194
Alimento      8.216879e-01 2.958241e-01  4.391940e-04 1.000000000
```

CUADRO X.2. *(Continuación)*^{STATISTICA}

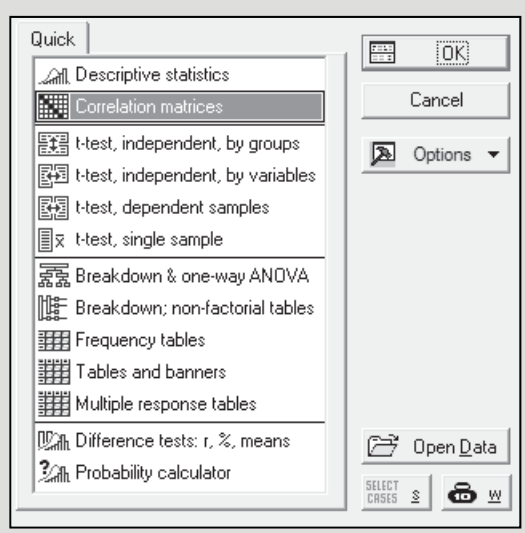
Aplicación con STATISTICA.

EJEMPLO. Utilizaremos los mismos datos del ejemplo del Cuadro X.1 en la aplicación con STATISTICA, los cuales se encuentran en el archivo **Cuadro X.1.sta**. Se observó que había una relación entre la longitud estándar y la longitud de la base de la aleta anal. Sin embargo, es posible que está relación sea diferente en función de la familia y/o el género y, para poder determinar las posibles diferencias asociadas a la familia o género al que pertenece la especie, es necesario analizar la correlación parcial. STATISTICA sólo calcula el coeficiente de Pearson.

Paso 1. Se selecciona «Statistics» ⇨ «Basic Statistics/Tables».

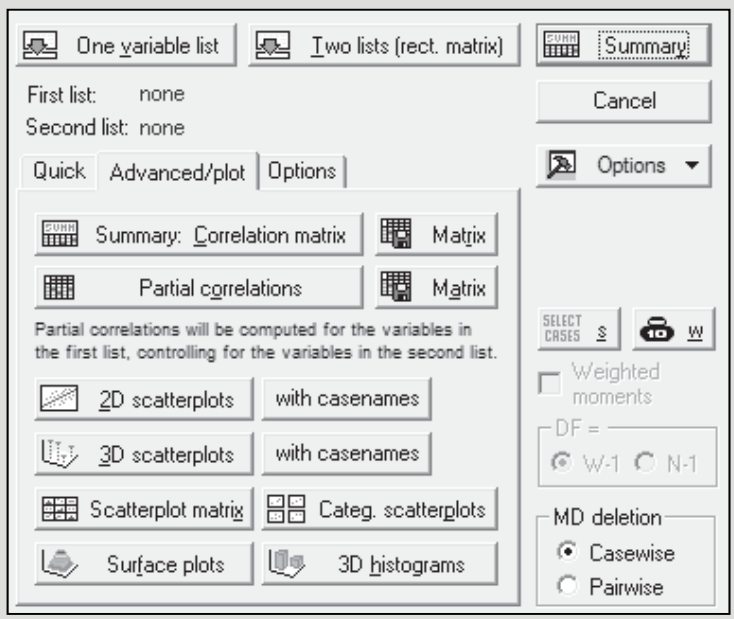


Paso 2. A continuación escogemos la opción «Correlation matrices».

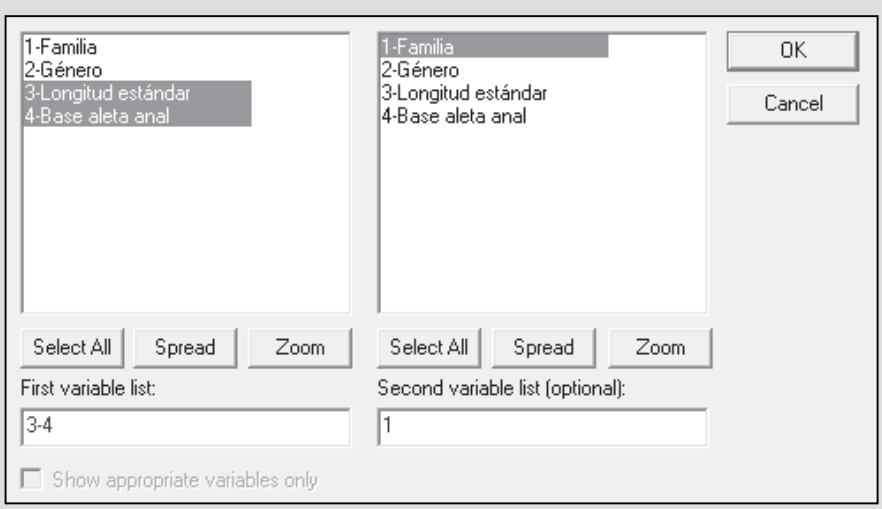


CUADRO X.2. *(Continuación)*[STATISTICA]

Paso 3. En el siguiente cuadro de diálogo debemos pulsar en el botón de «Two lists (rect. matrix)».



Paso 4. Aparece el siguiente cuadro de diálogo, en el cual debemos seleccionar las variables dependiente e independiente, que serán las mismas del ejemplo del Cuadro X.1, es decir, «Longitud estándar» y «Base aleta anal». Además, debemos seleccionar «Familia» como la variable cuyo posible efecto sobre las otras variables queremos controlar.

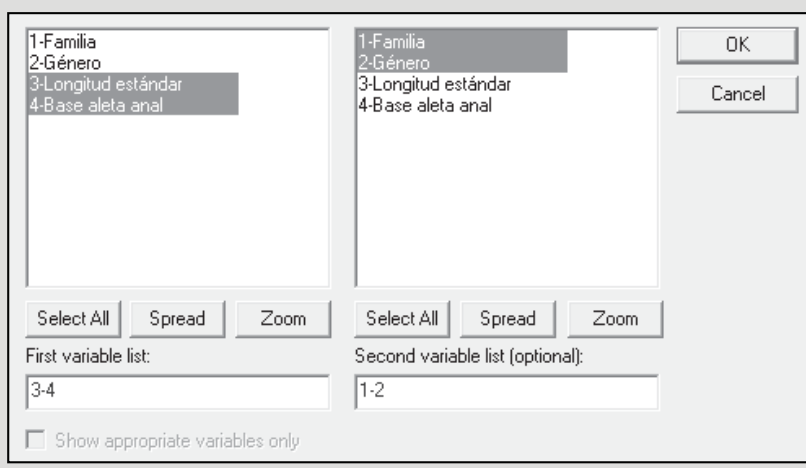


CUADRO X.2. *(Continuación)* STATISTICA

Paso 5. Pulsamos en «OK» y «Partial Correlations» en la ventana del paso 4 para ver los resultados. La relación entre las variables «Longitud estándar» y «Base aleta anal» varía muy poco dependiendo de la familia a la que pertenece la especie, ya que el valor del coeficiente de correlación de Pearson aumenta ligeramente (0,72), con respecto al valor del coeficiente de Pearson (0,71), cuando no se consideraba la familia a la que pertenecía la especie (véase Cuadro X.1).

Variable	Partial Correlations (Cuadro X.1.sta) Marked correlations are significant at p < ,05000 N=292 (Casewise deletion of missing data)	
	Longitud estándar	Base aleta anal
Longitud estándar	1,00	0,72
Base aleta anal	0,72	1,00

Paso 6. Por último, repetimos los pasos 3 y 4, pero añadiendo la variable «Género» además de «Familia».



Paso 7. Pulsando de nuevo en «Partial Correlations» en la ventana del paso 3, se observa en este caso que el coeficiente de correlación de Pearson sí aumenta significativamente (0,76), indicando que la relación entre «Longitud estándar» y «Base aleta anal» es diferente en función de la familia y el género al que pertenece la especie.

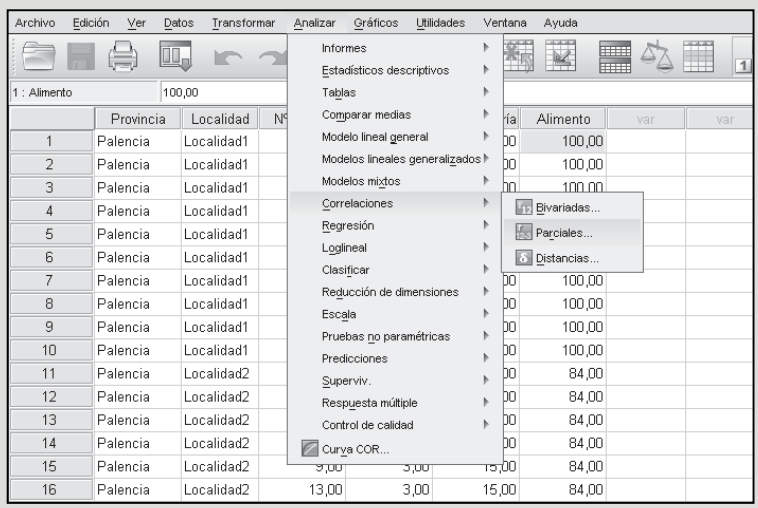
Variable	Partial Correlations (Cuadro X.1.sta) Marked correlations are significant at p < ,05000 N=292 (Casewise deletion of missing data)	
	Longitud estándar	Base aleta anal
Longitud estándar	1,00	0,76
Base aleta anal	0,76	1,00

CUADRO X.2. *(Continuación)*[SPSS]

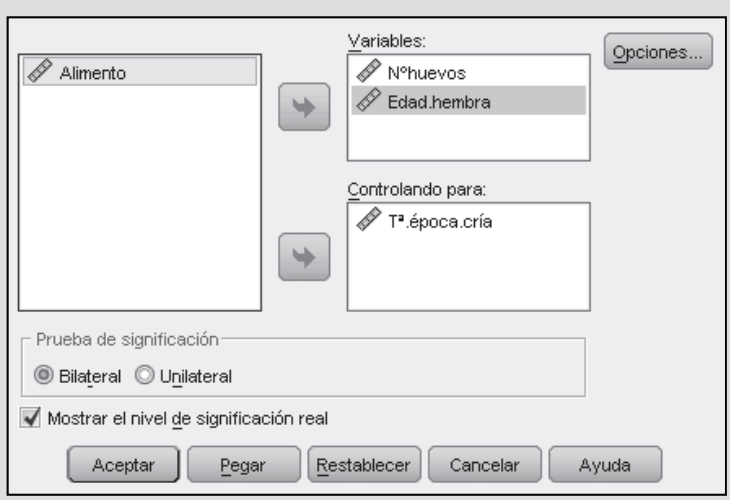
Aplicación con SPSS.

EJEMPLO. En el ejemplo del Cuadro X.1 en la aplicación con SPSS se observó que había una correlación positiva entre el número de huevos por hembra y la edad de las hembras. Sin embargo, esta relación puede variar en función del alimento y/o la temperatura durante la época de cría. Para determinar si efectivamente esta relación está afectada por otras variables es necesario analizar la correlación parcial.

Paso 1. Abrimos el archivo de datos **Cuadro X.1.sav** y vamos al menú «Analizar» ⇨ «Correlaciones» ⇨ «Parciales».



Paso 2. En el cuadro de diálogo que aparece a continuación, introducimos las variables «Nºhuevos» y «Edad.hembra» en el recuadro superior, y «Tª.época.cría» en el recuadro «Controlando para:».



CUADRO X.2. *(Continuación)*[SPSS]

Paso 3. Pulsamos en «Aceptar» y obtenemos el siguiente cuadro con los resultados.

Correlaciones

Variables de control			Nºhuevos	Edad.hembra
Tª.época.cría	Nºhuevos	Correlación	1,000	,842
		Significación (bilateral)	.	,000
		gl	0	97
	Edad.hembra	Correlación	,842	1,000
		Significación (bilateral)	,000	.
		gl	97	0

Se observa que el valor del coeficiente de Pearson es 0,84 y la relación es significativa ($p < 0{,}001$). En el ejemplo del Cuadro X.1, en el cual no se tenía en cuenta la temperatura en la época de cría, se observó que el coeficiente de Pearson era 0,64. Al ser ahora mayor el coeficiente, se concluye que la relación entre la edad de la hembra y el número de huevos varía en función de la temperatura durante la época de cría.

Paso 4. Repetimos el paso 2, pero utilizando las variables «Tª.época.cría» y «Alimento», para determinar si la cantidad de alimento también es una variable que puede estar influyendo. Se obtiene el siguiente cuadro de resultados.

Correlaciones

Variables de control			Nºhuevos	Edad.hembra
Tª.época.cría & Alimento	Nºhuevos	Correlación	1,000	,839
		Significación (bilateral)	.	,000
		gl	0	96
	Edad.hembra	Correlación	,839	1,000
		Significación (bilateral)	,000	.
		gl	96	0

Aunque se sigue rechazando la hipótesis de independencia, ya que $p < 0{,}001$, el valor del coeficiente de Pearson no cambió (0,84), con respecto al que obteníamos cuando solo se consideraba la variable «Tª.época.cría» (0,84). Por lo tanto, se concluye que la cantidad de alimento no afecta a la relación que existe entre la edad de la hembra y el número de huevos por hembra.

X.3. Correlación múltiple

X.3.1.COEFICIENTE DE CORRELACIÓN MÚLTIPLE DE PEARSON

Se trata de una extensión del método de correlación simple de Pearson a casos en los que tengamos más de una variable independiente. Este coeficiente

indica el grado de asociación entre varias variables independientes x_i y una variable dependiente *y*. Este coeficiente suele denotarse como R, y su cuadrado R^2 es el "coeficiente de determinación", que se interpreta como la proporción de varianza explicada por la regresión lineal múltiple:

$$R^2 = c' R_{xx}^{-1} c$$

donde *c* es una matriz o vector columna con todas las correlaciones posibles entre las variables independientes y la dependiente. *c'* es la traspuesta de esta matriz o vector fila. R_{xx} es una matriz de correlaciones entre todas las variables independientes, y el superíndice -1 indica que en la fórmula se utiliza la inversa de esta matriz. También se puede calcular como el coeficiente de correlación simple –bivariado– entre la variable dependiente *y* y la estimación de *y* obtenida mediante el modelo de regresión lineal múltiple, *y'*.

La significación se puede obtener por medio del estadístico *F*, calculado con la siguiente fórmula:

$$F = \frac{\frac{R^2_{yx_{1\ldots k}}}{k}}{\frac{(1 - R^2_{yx_{1\ldots k}})}{(n - k - 1)}}$$

donde el término $R^2_{yx_{1..k}}$ es el valor del coeficiente de correlación múltiple, *k es* el número de variables independientes, y *n* el número de datos. Los grados de libertad de la *F* son *k* y (*n* – *k* - 1).

X.3.2. COEFICIENTE DE CONCORDANCIA DE KENDALL (*W*)

Todos los coeficientes no paramétricos para medir asociación entre variables que hemos visto hasta ahora eran aplicables para datos bivariados. Por lo tanto, se basan en la comparación de dos grupos de rangos o de variables ordinales. El coeficiente de concordancia de Kendall (*W*), sin embargo, nos permitirá trabajar con varios grupos de rangos o variables ordinales. Es por tanto, en cierta medida, análogo al coeficiente de correlación múltiple de Pearson, aunque en este caso no existe una única variable dependiente. Supongamos que tenemos las puntuaciones de *m* jueces (por ejemplo catadores) sobre *n* objetos (por ejemplo vinos) y queremos medir el grado de acuerdo entre jueces. La fórmula para computarlo es la siguiente:

$$W = \frac{12S}{m^2(n^3 - n)}$$

donde:

$$S = \sum_{i=1}^{n} (R_i - \bar{R})^2$$

R_i es el rango medio para el objeto *i*, y $\bar{R}$ la media de todos los rangos:

$$R_i = \sum_{j=1}^{m} r_{i,j} \qquad \bar{R} = \frac{1}{2} m(n+1)$$

En caso de "empates" entre varios datos se asigna a cada dato la media de los rangos (si el tercero y el cuarto son iguales, su rango será 3,5 para ambos).

Para probar la significación de este estadístico tendríamos dos casos posibles:

- Si $3 \leq N \leq 7$ se emplea una distribución generada a partir de todas las combinaciones posibles de rangos, que existe tabulada para diferente número de variables o grupos ($3 \leq k \leq 20$), y amplio número de rangos (Siegel & Castellan 1988).
- Si $N > 7$ utilizaremos este otro estadístico, que sigue una distribución aproximadamente χ^2 con n - 1 grados de libertad:

$$\chi^2 = k(n-1)W$$

En cualquier caso, la H_0 que se establece será que el valor de W es igual a cero.

En este capítulo no se muestra un ejemplo del cálculo de estos coeficientes de correlación múltiple, puesto que los programas más habituales no disponen de funciones para su cálculo directo, sino dentro de un modelo de regresión múltiple, lo cual se verá en el siguiente capítulo.

Capítulo XI

REGRESIONES

En capítulos anteriores se describió la forma de determinar si existía una relación significativa entre dos variables cualitativas (Capítulo VIII) y cuantitativas (Capítulo X), es decir, si las variables eran dependientes o independientes. En este capítulo vamos a analizar cómo describir el tipo de función que mejor se ajusta a la posible relación que existe entre variables.

Se define *regresión* como la teoría que trata de expresar mediante una función matemática la relación que existe entre una variable dependiente y una (regresión simple) o varias (regresión múltiple) variables independientes. La obtención de esta función permite predecir cual será el valor de la variable dependiente en función del valor que tome la variable o variables independientes.

La regresión se diferencia de la *correlación* en que esta última estudia el grado de asociación entre las variables, y determina si la relación es o no es significativa, mientras que la regresión, como se mencionó anteriormente, trata de definir la función que mejor explica la relación entre las variables.

En este capítulo vamos a estudiar los modelos de regresión que existen cuando la variable dependiente es cuantitativa y los que hay para variables dependientes cualitativas.

XI.1. Modelos de regresión para variables dependientes cuantitativas

XI.1.1. REQUISITOS

Para aplicar un modelo de regresión entre variables no se requiere que los datos presenten una distribución normal o que exista homogeneidad de varianzas; sin embargo, para poder determinar si la función obtenida con el modelo de regresión es significativa es necesario aplicar contrastes, y para ello se tienen que cumplir los siguientes requisitos, que en gran medida están relacionados con los residuos (diferencia entre el valor observado de la variable dependiente y el valor ajustado por la función):

1. Los residuos obtenidos del modelo de regresión deben presentar una distribución Normal.
2. Debe existir homocedasticidad en los residuos, es decir, la varianza de los mismos debe ser constante. Las gráficas que se muestran a continuación (Figura XI.1) son ejemplos en los que no se cumple la homocedasticidad de los residuos, porque su variabilidad no es constante, sino que cambia dependiendo de los valores pronosticados. En la primera gráfica, obtenida en el ajuste de una recta, se observa además, que la relación no es lineal, ya que no se mantiene la tendencia horizontal.

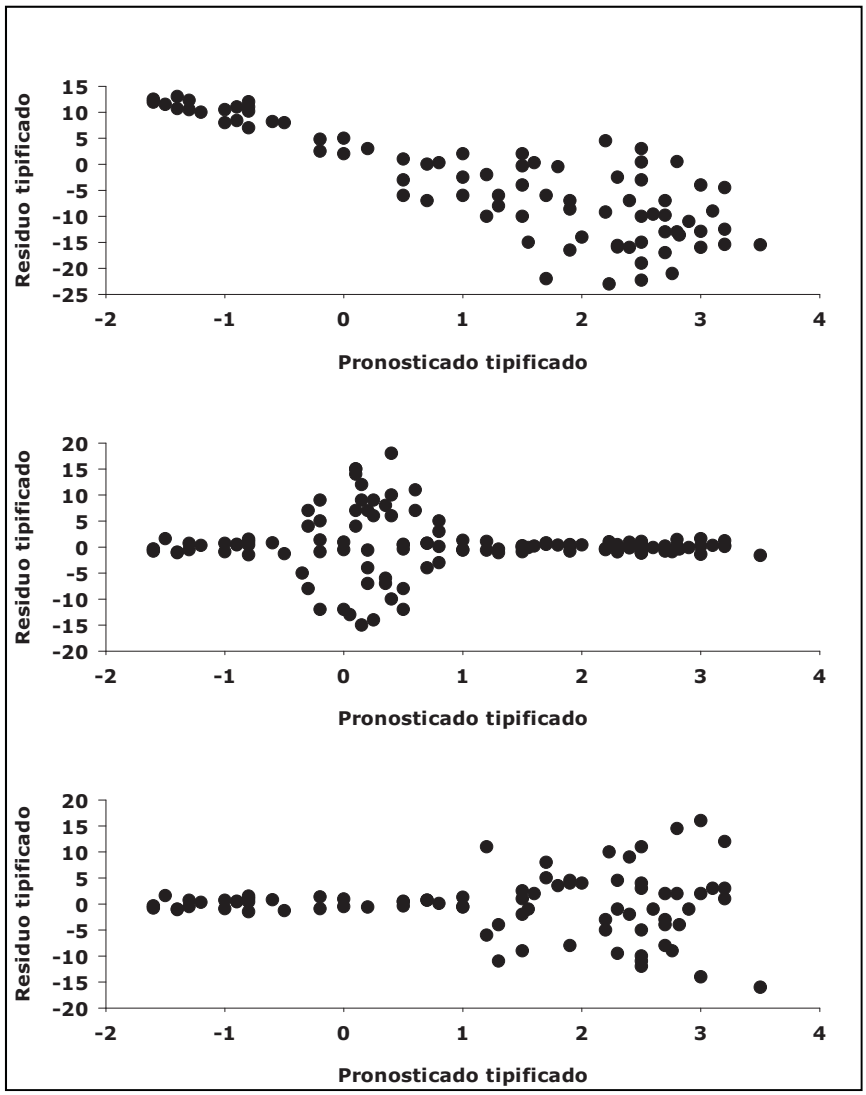


Figura XI.1. Ejemplos en los que no se cumple la homocedasticidad.

3. No debe existir autocorrelación en la serie de residuos (deben ser independientes). Si este requisito no se cumple, no es posible saber el grado de relación exacta entre la variable independiente y la dependiente, ya que parte de la predicción se debe a los propios valores de la variable dependiente.
4. En el caso del modelo de regresión múltiple, no debe existir relación entre las variables independientes, es decir, no debe existir multicolinealidad.

XI.1.2. REGRESIÓN SIMPLE

Lo primero que hay que hacer cuando se intenta buscar el mejor ajuste entre dos variables, es representar los datos, poniendo en el eje *Y* la variable dependiente y en el eje *X* la independiente. Esto es muy importante, ya que, como se mencionó en el Capítulo I, la representación de los datos es necesaria para ver el tipo de relación que existe entre dos variables y para identificar posibles *outliers*.

Las funciones más comunes son las que se muestran en la Figura XI.2 y se expresan a continuación:

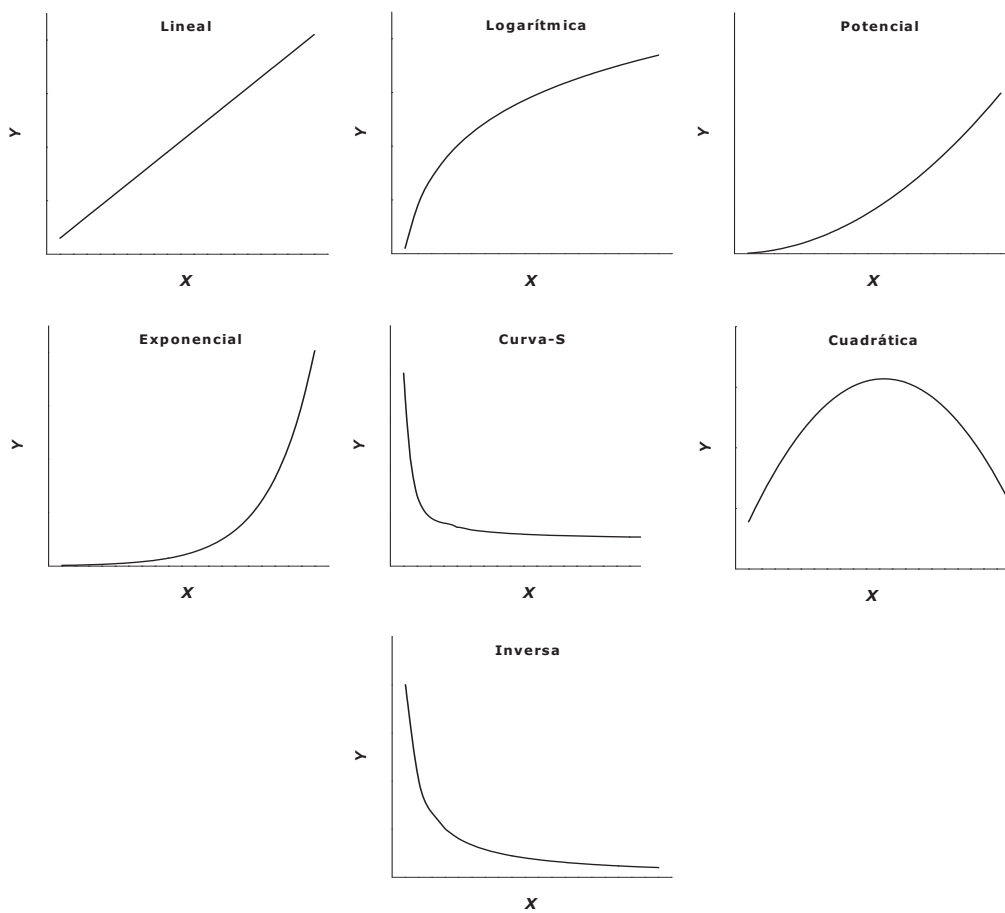


Figura XI.2. Representación gráfica de las funciones mostradas en el texto.

Lineal	$y = a + bx$		
Logarítmica	$y = a + b\ln x$		
Potencial	$y = ax^b$	ó	$\ln y = \ln a + b\ln x$
Exponencial	$y = ae^{bx}$	ó	$\ln y = \ln a + bx$
Curva-S	$y = e^{a+\frac{b}{x}}$	ó	$\ln y = a + \frac{b}{x}$
Cuadrática o Parábola	$y = a + bx + cx^2$		
Inversa	$y = a + \frac{b}{x}$		

CUADRO XI.1. Regresión simple

Aplicación con R.

EJEMPLO. En el archivo **Cuadro XI.1.V.csv** hay datos del peso (gr), longitud (cm), anchura (cm) y altura (cm) de la concha de tres especies de moluscos. El objetivo es determinar cual es el mejor modelo de regresión que existe entre el peso (variable dependiente) y el ancho de la concha (variable independiente).

Después de invocar al programa R y seleccionar el directorio de trabajo como se explica en el Apéndice II, debemos abrir el *script* **Cuadro XI.1.R** que se muestra en la siguiente ventana. Para este *script* es necesario tener instalados, siguiendo las instrucciones del Apéndice II, los paquetes "nortest", "car" y "lmtest".

```
# Cuadro XI.1
# Regresión lineal simple

#########################################################
# Sección modificable por el usuario
#########################################################
datos<-read.csv2("Cuadro XI.1.V.csv",enc="latin1")

#Variable dependiente
varRespuesta<-"Peso.gr"

#Variable independiente
varExplicativa<-"Ancho.cm"

# Modelos posibles:
# 1. Lineal
# 2. Logarítmico
# 3. Potencial
# 4. Exponencial
# 5. Curva-S
# 6. Cuadrática o Parábola
# 7. Inversa
```

CUADRO XI.1. *(Continuación)*[R]

Paso 1. En primer lugar es necesario importar los datos desde nuestro archivo csv.

```
datos<-read.csv2("Cuadro XI.1.V.csv",header=T,encoding="latin1")
```

Paso 2. Se seleccionan las variables dependiente e independiente.

```
#Variable dependiente
varRespuesta<-"Peso.gr"

#Variable independiente
varExplicativa<-"Ancho.cm"
```

Paso 3. Se seleccionan los modelos de regresión que se quieren estimar. En este ejemplo se escogen todos los modelos posibles.

```
# 1. Lineal
# 2. Logarítmico
# 3. Potencial
# 4. Exponencial
# 5. Curva-S
# 6. Cuadrática o Parábola
# 7. Inversa

Selecmodelos<-c(1,2,3,4,5,6,7)
#Selecmodelos<-c(1,3,4)
```

Paso 4. También es posible seleccionar dos grupos en dos variables diferentes. En este ejemplo se hacen los modelos de regresión sólo para la especie 3.

```
varSeleccion1<-list(c("Especie","Molusco 3"))
#varSeleccion1<-NULL
#varSeleccion2<-list(c("Especie","Molusco 1","Molusco 2"))
varSeleccion2<-NULL
```

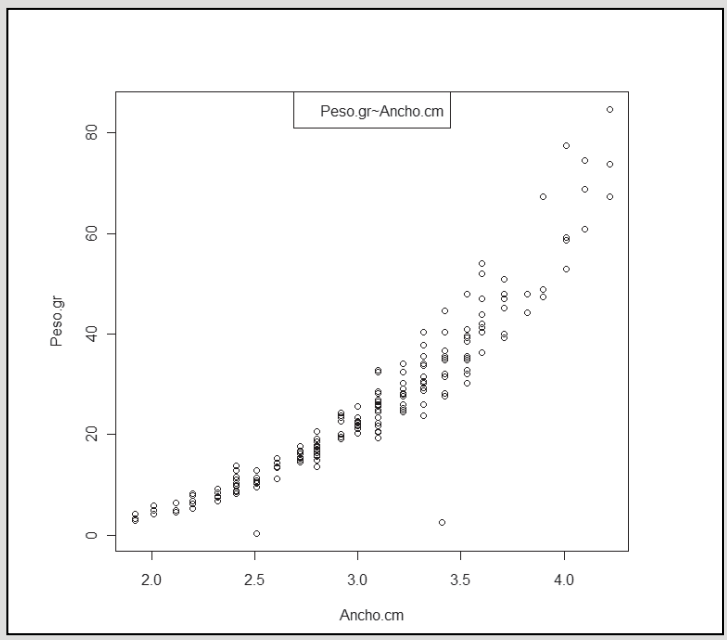
Paso 5. Se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto, y también el archivo donde se guardarán los residuos, que tiene formato csv y se puede abrir con Excel.

```
nomSalida<-"Salida Cuadro XI.1.txt"
CSV<-"Salida Cuadro XI.1.csv"
```

Paso 6. Por último, seleccionamos el contenido de este *script* pulsando *ctrl+A* y lo ejecutamos pulsando *ctrl+R*. Es necesario pulsar varias veces consecutivas con el ratón sobre la gráfica que sale, para que salgan todas las gráficas de los diferentes modelos de regresión. Para poder ver todas las gráficas hay que arrastrar cada una de las ventanas, y así poder ver la gráfica que hay debajo.

CUADRO XI.1. *(Continuación)*[R]

Paso 7. El gráfico muestra como la relación parece que no es lineal y que hay 2 datos, que posiblemente son *outliers*, los cuales conviene identificar –y tal vez eliminar– antes de proseguir con el análisis.



En el archivo **Salida Cuadro XI.1.txt** se observa que, en ninguno de los modelos, se cumple el requisito de que los residuos tengan una distribución Normal, ya que en todos $p < 0{,}05$, tanto en el test de Shapiro-Wilk como en el de Kolmogorov-Smirnov. El modelo se puede identificar por el número que hay en el código, es decir, «residuos$”1”» corresponde al primer modelo que es el de la regresión lineal.

```
[[1]]

        Shapiro-Wilk normality test

data:  residuos$"1"
W = 0.881, p-value = 3.184e-11


[[2]]

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos$"1"
D = 0.1224, p-value = 1.935e-07


[[1]]

        Shapiro-Wilk normality test

data:  residuos$"2"
W = 0.8732, p-value = 1.186e-11


[[2]]

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos$"2"
D = 0.1276, p-value = 4.099e-08
```

CUADRO XI.1. *(Continuación)*[R]

Paso 8. En el archivo **Salida Cuadro XI.1.csv** se pueden identificar los dos datos anteriores que posiblemente sean *outliers*, ya que son los que tienen unos residuos muy altos (señalados en negrita).

	datos...varRespuesta.	residuos..1.	residuos..2.	residuos..3.	residuos..4.	residuos..5.	residuos..6.	residuos..7.
1	32,14	-4,606562457	-4,663335546	-0,062382418	-0,040618886	-0,064884242	-3,111405438	-4,3044086
2	13,33	-0,670326645	-1,847655005	0,070967709	0,103545848	0,015763357	0,52874163	-3,416014473
3	46,99	2,099723487	3,6746194	0,012280053	-0,027491851	0,07296794	0,175541488	5,584121062
4	28,5	0,739604778	-0,443368401	0,185758346	0,243825406	0,13447373	3,850623162	-1,392397937
5	28,79	-5,148385196	-5,639004646	-0,061188493	-0,024241653	-0,08171378	-2,905144183	-5,742583199
6	34,81	-1,936562457	-1,993335546	0,017420923	0,039184456	0,0149191	-0,441405438	-1,6344086
7	47,94	-0,039271499	2,286870424	-0,07725739	-0,146571973	0,010808456	-3,746698795	4,849245505
8	14,88	-2,209321632	-3,600562549	0,026185962	0,074451206	-0,038282277	-0,119098164	-5,229547457
9	28,22	0,459604778	-0,723368401	0,175885205	0,233952265	0,124600589	3,570623162	-1,672397937
10	35,52	-1,226562457	-1,283335546	0,03761214	0,059375672	0,035110316	0,268594562	-0,9244086
11	24,29	1,584323847	0,132654351	0,250204192	0,311596941	0,185163291	4,609652498	-1,285812144
12	26,84	-0,920395222	-2,103368401	0,125747571	0,18381463	0,074462955	2,190623162	-3,052397937
13	17,12	-2,21586344	-3,679828698	0,057728597	0,113520156	-0,009263699	0,354494021	-5,269760726
14	0,25	**-10,94214938**	**-11,80190283**	**-3,758863095**	**-3,746315158**	**-3,79893765**	**-10,78529827**	**-13,18242892**
15	20,7	1,36413656	-0,099828698	0,247614927	0,303406485	0,18062263	3,934494021	-1,689760726
16	15,64	-3,69586344	-5,159828698	-0,032687039	0,02310452	-0,099679335	-1,125505979	-6,749760726
17	15,33	-1,759321632	-3,150562549	0,055979626	0,104244869	-0,008488613	0,330901836	-4,779547457
18	12,81	4,426027877	4,010946769	0,330094181	0,316655808	0,312604974	3,319636998	2,96614299
19	22,42	-2,532217962	-3,899884853	0,068750783	0,130325225	0,008364616	0,619604949	-5,138244138
20	8,33	5,843200124	6,825325532	0,24156465	0,151839175	0,301579468	1,364565225	7,084035048
21	40,85	1,014442557	1,513791425	0,058726175	0,060091037	0,078454914	1,431423499	2,427691581
22	20,09	-2,615676153	-4,067345649	0,060361628	0,121754377	-0,004679273	0,409652498	-5,485812144
23	9,87	1,486027877	1,070946769	0,069367919	0,055929545	0,051878712	0,379636998	0,02614299
24	32,79	5,029604778	3,846631599	0,325977849	0,384044909	0,274693234	8,140623162	2,897602063
25	14,59	-2,499321632	-3,890562549	0,006504295	0,054769539	-0,057963944	-0,409098164	-5,519547457
26	44,7	7,953437543	7,896664454	0,267489723	0,289253256	0,264987899	9,448594562	8,2555914
27	32,46	4,699604778	3,516631599	0,315862821	0,37392988	0,264578205	7,810623162	2,567602063
28	26,58	-1,180395222	-2,363368401	0,116013312	0,174080372	0,064728696	1,930623162	-3,312397937
29	27,66	-3,470207935	-4,322053088	0,013441097	0,062168058	-0,022719305	-0,699907762	-4,842010879
30	30,57	-3,368385196	-3,859004646	-0,001197462	0,035749379	-0,021722748	-1,125144183	-3,962583199
31	32,14	-7,695557443	-7,196208575	-0,181079286	-0,179714424	-0,161350547	-7,278576501	-6,282308419
32	35,52	1,581614804	1,090995354	0,14887932	0,185826161	0,128354034	3,824855817	0,987416801
33	**2,55**	**-33,91574473**	**-34,01904818**	**-2,585411118**	**-2,561981838**	**-2,589813969**	**-32,33583319**	**-33,70827193**
34	47,94	8,104442557	8,603791425	0,218769583	0,220134445	0,238498321	8,521423499	9,517691581

Paso 9. Al realizar de nuevo el análisis eliminando esos 2 datos se observa que la regresión que mejor se ajusta a los datos es la potencial (modelo 3), con un r^2 = 0,96, donde la pendiente es significativa (p < 0,001).

```
Modelo: log(Peso.gr)~log(Ancho.cm)
Tabla ANOVA
Analysis of Variance Table

Response: log(Peso.gr)
              Df Sum Sq Mean Sq F value    Pr(>F)
log(Ancho.cm)   1 86.599  86.599    5031 < 2.2e-16 ***
Residuals     189  3.253   0.017
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Estimación de parámetros

Call:
lm(formula = formula, data = datos)

Residuals:
      Min        1Q    Median        3Q       Max
-0.346725 -0.096505  0.004282  0.081015  0.355159

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.98560    0.05804  -16.98   <2e-16 ***
log(Ancho.cm)   3.69561    0.05210   70.93   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1312 on 189 degrees of freedom
Multiple R-squared: 0.9638,     Adjusted R-squared: 0.9636
F-statistic:  5031 on 1 and 189 DF,  p-value: < 2.2e-16
```

CUADRO XI.1. *(Continuación)*[R]

Paso 10. Normalidad de los residuos. Los residuos tienen una distribución Normal, ya que tanto en el test de Shapiro-Wilk como en el de Kolmogorov-Smirnov $p > 0,05$.

```
[[1]]

        Shapiro-Wilk normality test

data:  residuos$"3"
W = 0.9942, p-value = 0.6575


[[2]]

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuos$"3"
D = 0.0482, p-value = 0.3418
```

Paso 11. Autocorrelación. No se cumple el requisito de que no debe existir autocorrelación ya que en la prueba de Durbin-Watson $p < 0,05$, aunque está muy próximo a 0,05 ($p = 0,046$).

```
$`3`
 lag Autocorrelation D-W Statistic p-value
   1       0.1214283      1.735356   0.046
 Alternative hypothesis: rho != 0
```

El hecho de que exista una pequeña autocorrelación en los residuos solo significa que probablemente el modelo se puede mejorar añadiendo algún regresor más, pero no lo invalida.

Paso 12. Homocedasticidad. Por último, si se cumple el requisito de homocedasticidad de los residuos, ya que la probabilidad del test Breusch-Pagan es mayor que 0,05.

```
$`3`

        studentized Breusch-Pagan test

data:  modelo
BP = 1.6576, df = 1, p-value = 0.1979
```

El hecho de que no se cumplan los 3 requisitos no invalida el modelo ya que es muy predictivo, el 96,38% de la variación observada en el peso (P en gr) se puede explicar en función del ancho de la concha (A en cm):

$$\ln(P) = -0,9856 + 3,695*\ln(A)$$

o en su otro formato

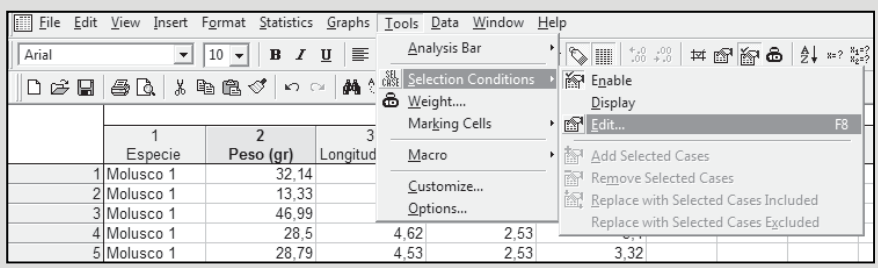
$$P = e^{-0,9856}*A^{3,695}$$
$$P = 0,3732*A^{3,695}$$

CUADRO XI.1. *(Continuación)*[STATISTICA]

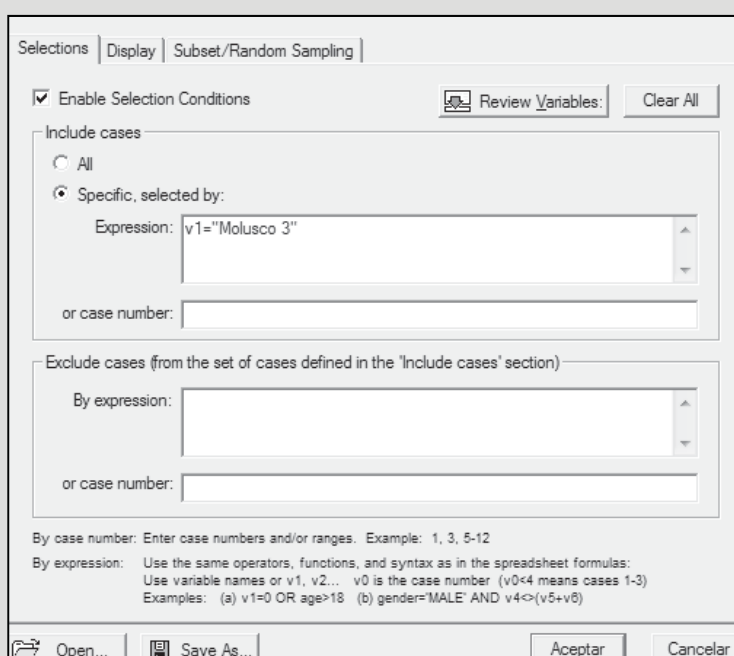
Aplicación son STATISTICA.

EJEMPLO. Se utilizarán los mismos datos que en la aplicación con R, que están en el archivo **Cuadro XI.1.sta**. Son datos del peso (gr), longitud (cm), anchura (cm) y altura (cm) de la concha de tres especies de moluscos. En este caso el objetivo es determinar cual es el mejor modelo de regresión que existe entre el peso (variable dependiente) y la longitud de la concha (variable independiente) para la especie 3.

Paso 1. Debido a que existen varias especies, es necesario primero definir la especie con la que se va a trabajar. Para seleccionar los casos vamos a «Tools» ⇨ «Selection Conditions» ⇨ «Edit…».

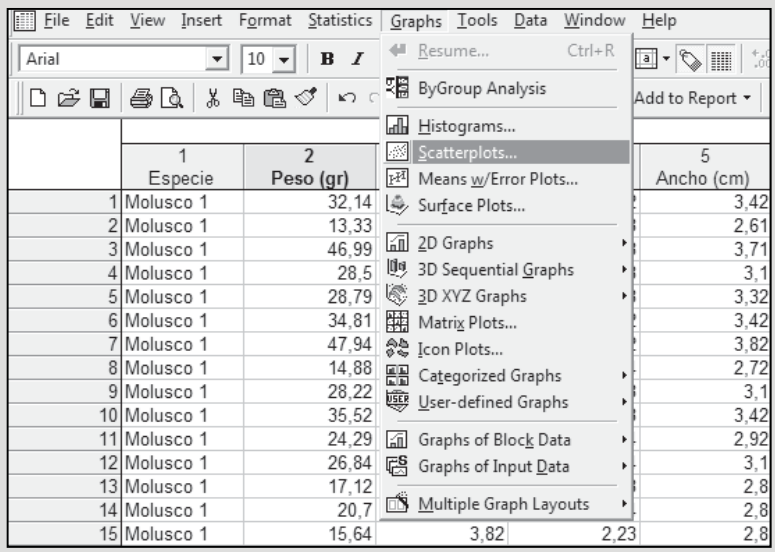


Paso 2. En la pestaña «Selections» se marca «Enable Selection Conditions» ⇨ «Specific, selected by:» donde se selecciona la variable o variables a analizar, usando OR y AND si se quieren combinar varias condiciones. En este caso se selecciona «v1="Molusco 3"».

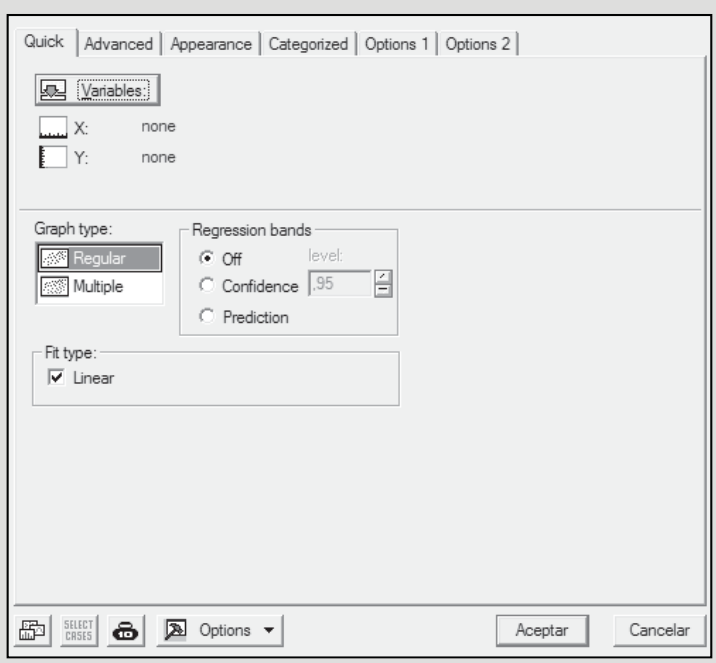

Selections
Display
Subset/Random Sampling
Enable Selection Conditions
Review Variables:
Clear All
Include cases
All
Specific, selected by:
Expression: v1="Molusco 3"
or case number:
Exclude cases (from the set of cases defined in the 'Include cases' section)
By expression:
or case number:
By case number: Enter case numbers and/or ranges. Example: 1, 3, 5-12
By expression: Use the same operators, functions, and syntax as in the spreadsheet formulas:
Use variable names or v1, v2... v0 is the case number (v0<4 means cases 1-3)
Examples: (a) v1=0 OR age>18 (b) gender='MALE' AND v4<>(v5+v6)
Open...
Save As...
Aceptar
Cancelar

CUADRO XI.1. *(Continuación)* *STATISTICA*

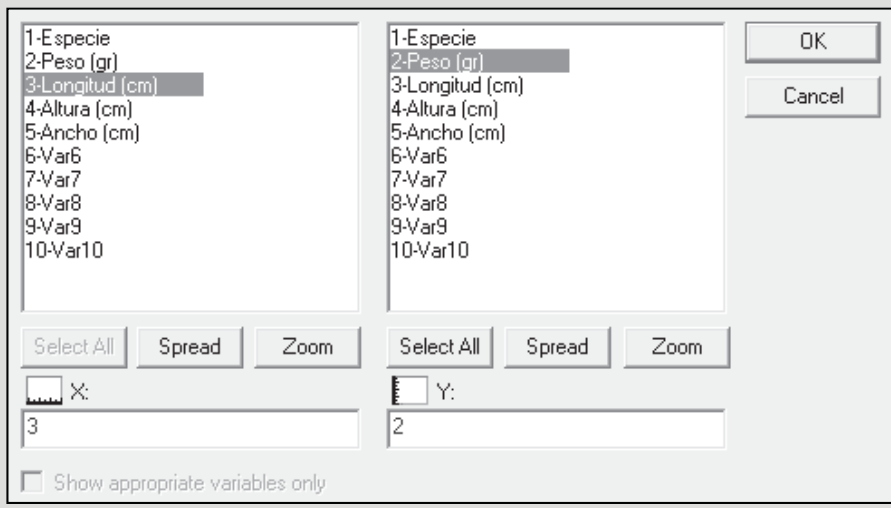
Paso 3. A continuación es importante graficar los resultados para ver el tipo de relación que existe entre las variables y si hay posibles *outliers*. Para ello vamos a «Graphs» ⇨ «Scatterplots…».



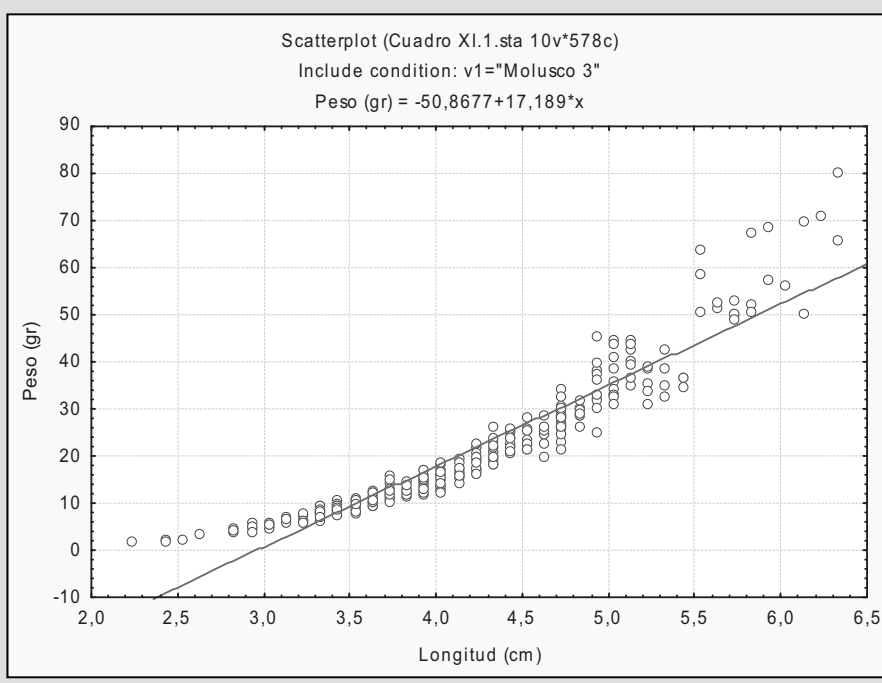
Paso 4. Al pulsar en «Variables» aparece la ventana que se muestra posteriormente donde seleccionamos las variables «Peso (gr)» y «Longitud (cm)».



CUADRO XI.1. *(Continuación)*[STATISTICA]

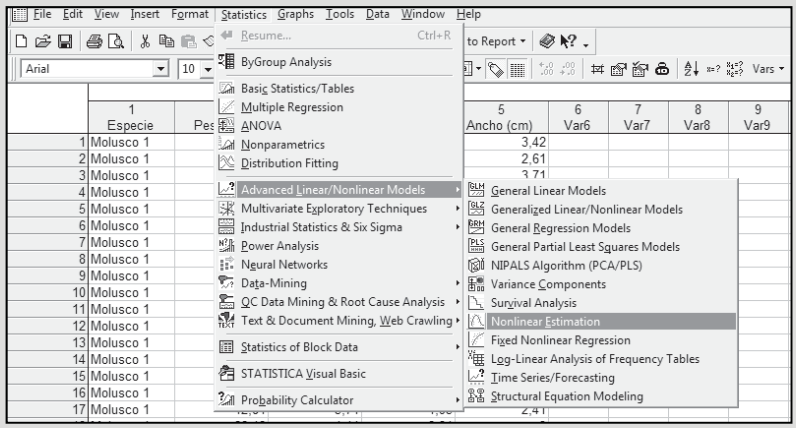


Paso 5. Se observa una relación que parece que no es lineal.

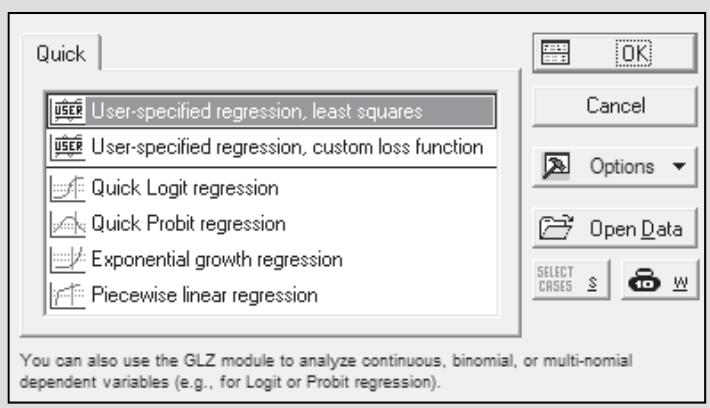


Paso 6. Para calcular los modelos de regresión se selecciona «Statistics» ⇨«Advanced Linear/Nonlinear Models» ⇨ «Nonlinear Estimation».

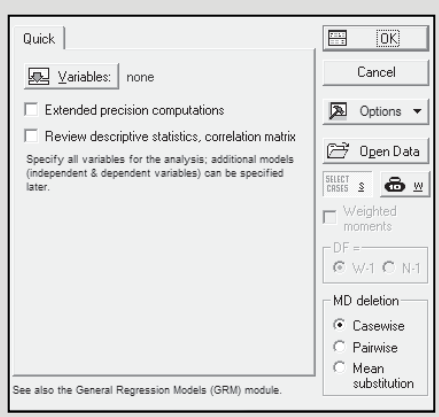
CUADRO XI.1. *(Continuación)* *STATISTICA*



Paso 7. En la ventana siguiente se selecciona «User-specified regression, least squares».

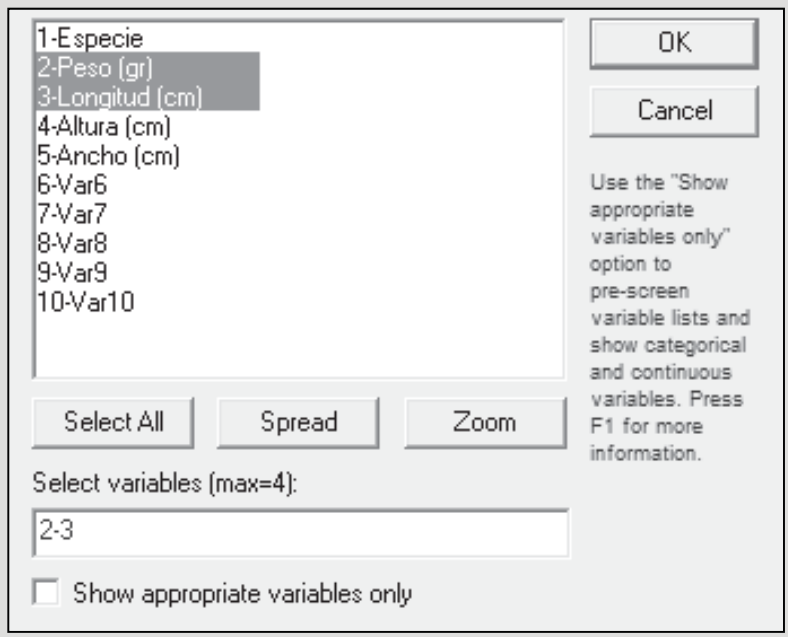


Paso 8. A continuación se pulsa en «Variables».

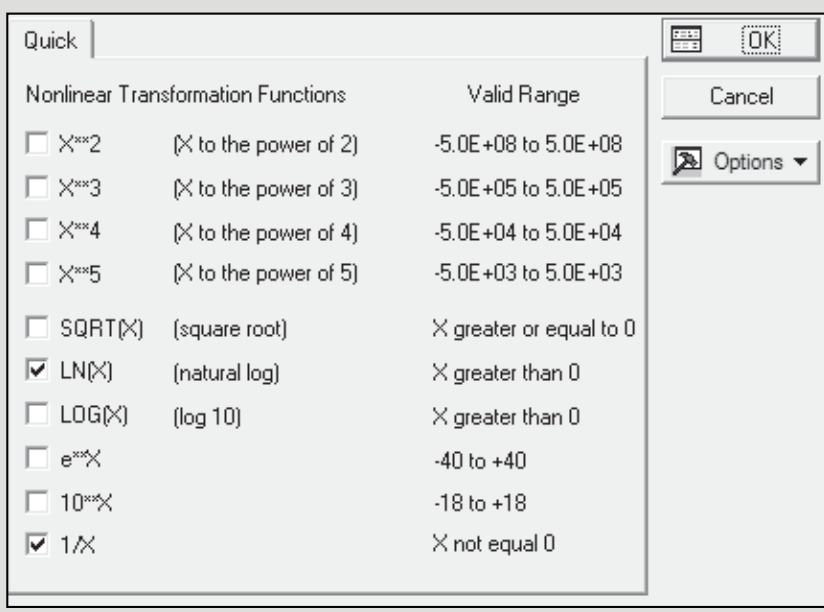


CUADRO XI.1. *(Continuación)*[STATISTICA]

Paso 9. Se seleccionan las dos variables de interés: «Peso (gr)» y «Longitud (cm)».

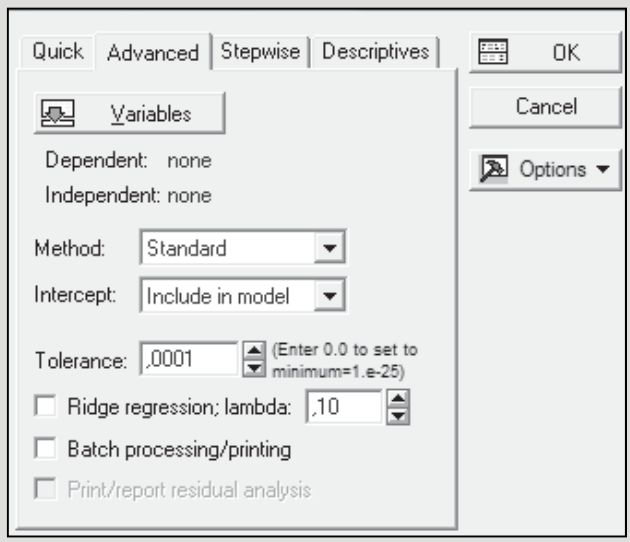


Paso 10. A continuación se marca «LN(X)» que es necesario para casi todos los modelos de regresión y «1/X» que hace falta para la Curva-S.

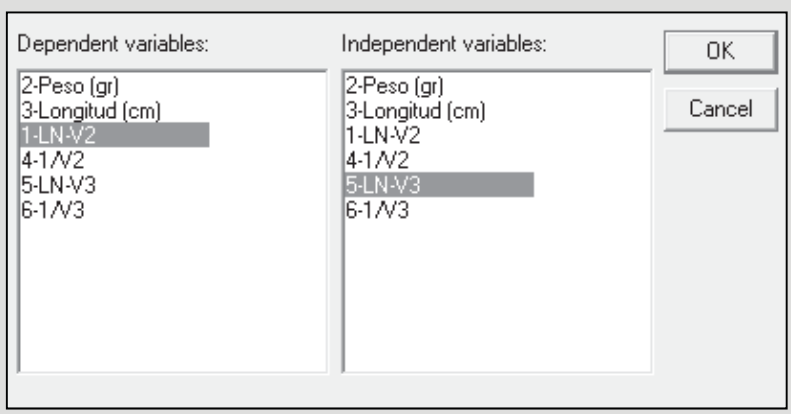

Quick
OK
Cancel
Options
Nonlinear Transformation Functions
Valid Range
X**2 (X to the power of 2) -5.0E+08 to 5.0E+08
X**3 (X to the power of 3) -5.0E+05 to 5.0E+05
X**4 (X to the power of 4) -5.0E+04 to 5.0E+04
X**5 (X to the power of 5) -5.0E+03 to 5.0E+03
SQRT(X) (square root) X greater or equal to 0
LN(X) (natural log) X greater than 0
LOG(X) (log 10) X greater than 0
e**X -40 to +40
10**X -18 to +18
1/X X not equal 0

CUADRO XI.1. *(Continuación)*[STATISTICA]

Paso 11. En el siguiente cuadro en la pestaña «Advanced» se pulsa en «Variables».



Paso 12. A continuación se seleccionan la variable dependiente y la variable independiente en función del modelo de regresión que se quiere calcular, en base a lo mostrado de la Figura XI.2. En la siguiente ventana se muestra el ejemplo para el modelo Potencial, para el cual es necesario seleccionar el logaritmo de ambas variables; se seleccionaría variable «LN-V2» como dependiente (peso) y la independiente sería «LN-V3» (longitud).

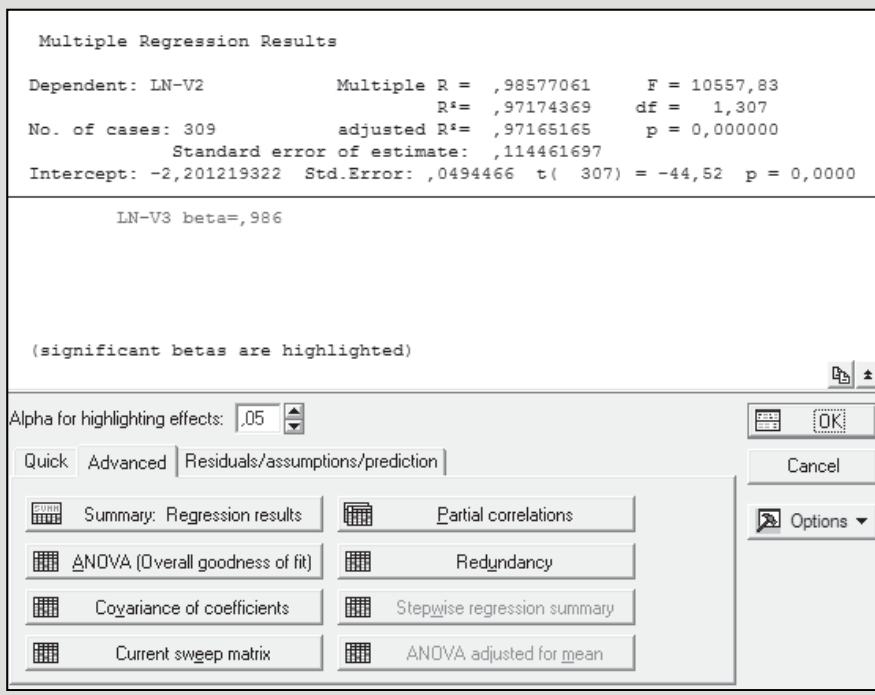


Si fuera el modelo Curva-S se seleccionaría «LN-V2» (variable dependiente) y «1/V3» (variable independiente). Para el modelo Lineal se escogería «Peso (gr)» (variable dependiente) y «Longitud (cm)» (variable independiente). Para el modelo Exponencial «LN-V2» (variable dependiente) y «Longitud (cm)» (variable independiente).

CUADRO XI.1. *(Continuación)*[STATISTICA]

Para el modelo Logarítmico «Peso (gr)» (variable dependiente) y «LN-V3» (variable independiente).

Paso 13. Al probar los diferentes modelos se observa que el que mejor se ajusta es el Potencial, con un r^2 = 0,9717.



Paso 14. En la ventana anterior, en la pestaña «Advanced», al pulsar en «Summary: Regression results», se obtienen los coeficientes de la regresión.

	Regression Summary for Dependent Variable: LN-V2 (Cuadro XI.1.sta) R= ,98577061 R²= ,97174369 Adjusted R²= ,97165165 F(1,307)=10558, p<0,0000 Std.Error of estimate: ,11446 Include condition: v1="Molusco 3"					
N=309	Beta	Std.Err. of Beta	B	Std.Err. of B	t(307)	p-level
Intercept			-2,20122	0,049447	-44,5171	0,00
LN-V3	0,985771	0,009594	3,55898	0,034637	102,7513	0,00

El peso (*P* en gr) se relaciona con la longitud de la concha (*L* en cm) por medio de la siguiente ecuación:

$$\ln(P) = -2{,}201 + 3{,}5589 * \ln(L)$$

CUADRO XI.1. *(Continuación)* STATISTICA

o en su otro formato

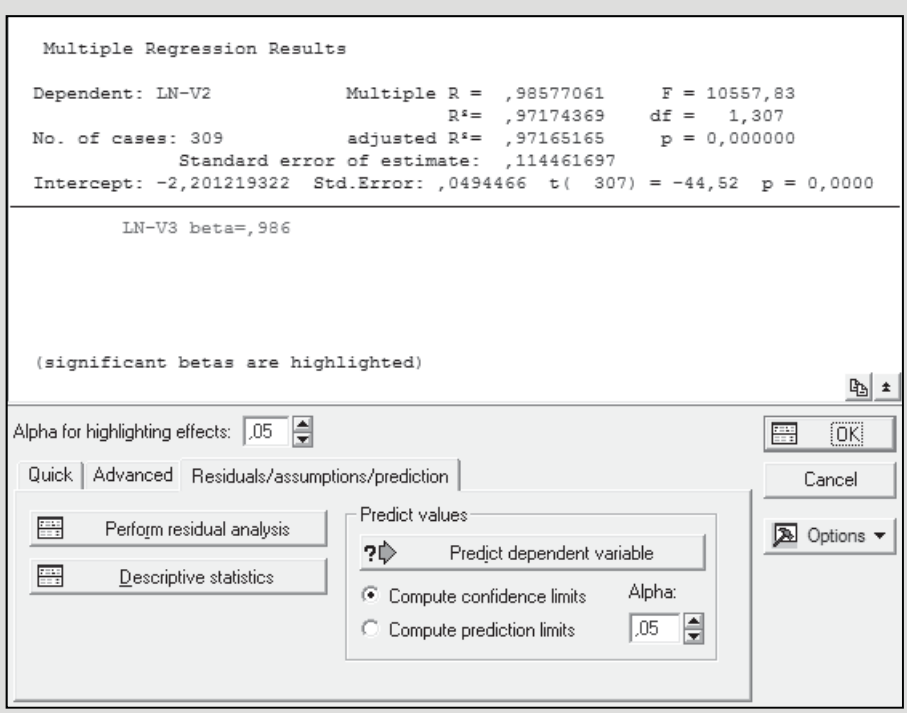
$$P = e^{-2,201} * L^{3,5589}$$

$$P = 0,1106 * L^{3,5589}$$

Paso 15. En la ventana del paso 13, al pulsar en «ANOVA (Overall goodness of fit)», se obtiene la tabla de ANOVA. Por tanto, existe una relación significativa entre el peso y la longitud de la concha ($F_{1,307}$ = 10557, p < 0,001, r^2 = 0,97).

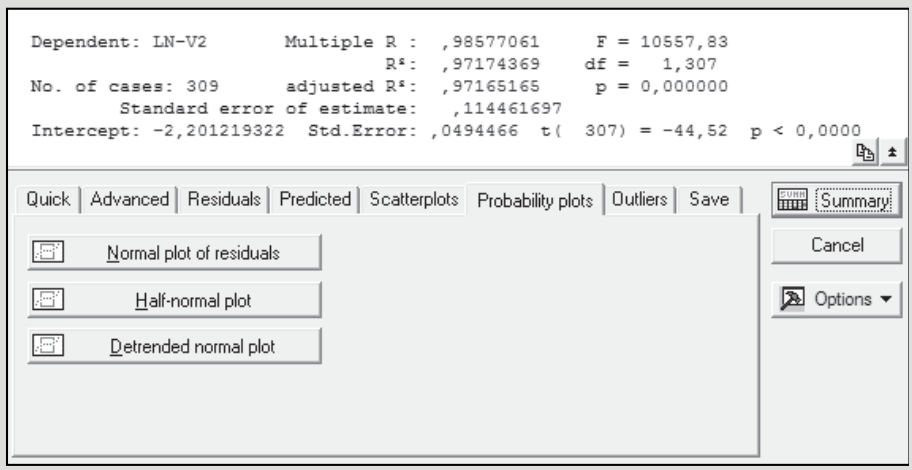
	Analysis of Variance; DV: LN-V2 (Cuadro XI.1.sta) Include condition: v1="Molusco 3"				
Effect	Sums of Squares	df	Mean Squares	F	p-level
Regress.	138,3232	1	138,3232	10557,83	0,00
Residual	4,0222	307	0,0131		
Total	142,3453				

Paso 16. Para comprobar si se cumplen los tres requisitos, en la pestaña «Residuals/assumptions/prediction» se pulsa en «Perform residual analysis».

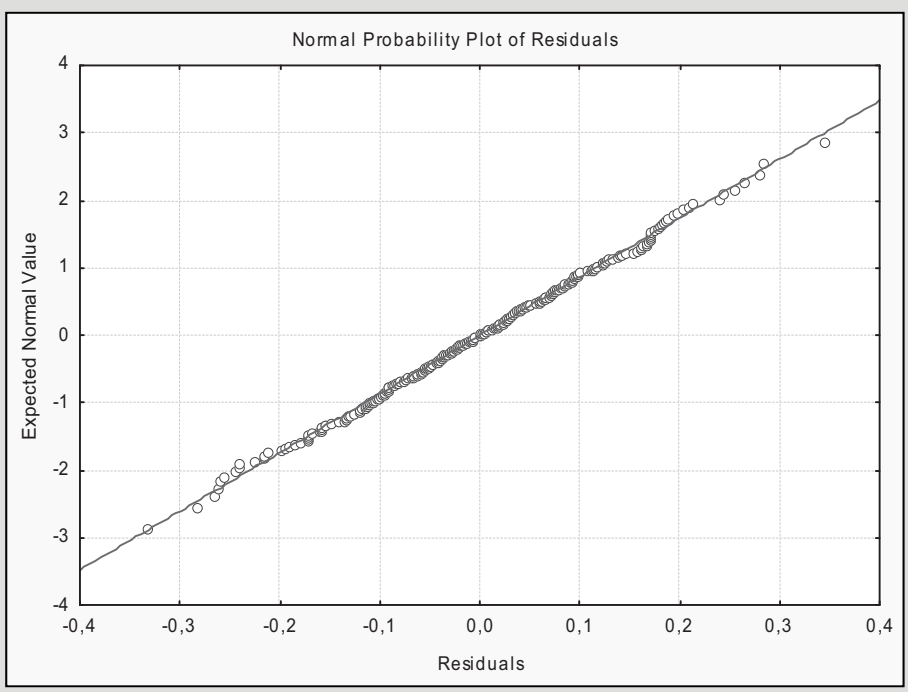


CUADRO XI.1. *(Continuación)*[STATISTICA]

Paso 17. Normalidad de los residuos. En la ventana que aparece se pulsa en «Normal plot of residuals».

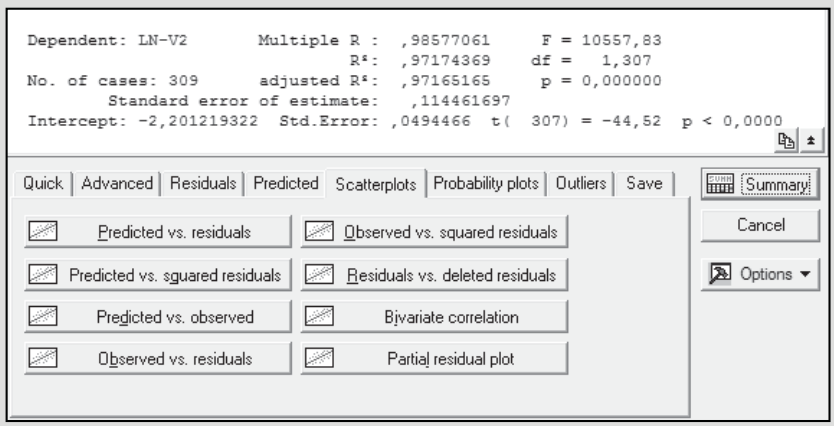


El gráfico que se obtiene muestra que los residuos de ajustan bien a una distribución Normal, ya que todos los puntos quedan muy cercanos a la recta.

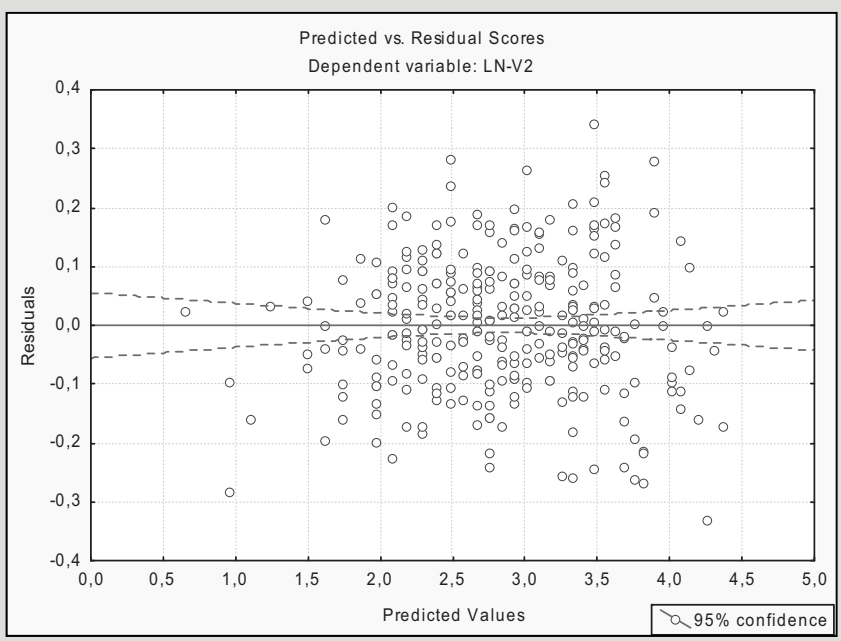


CUADRO XI.1. *(Continuación)* *STATISTICA*

Paso 18. Homocedasticidad de los residuos. En la pestaña «Scatterplots» se pulsa en «Predicted vs. residuals».

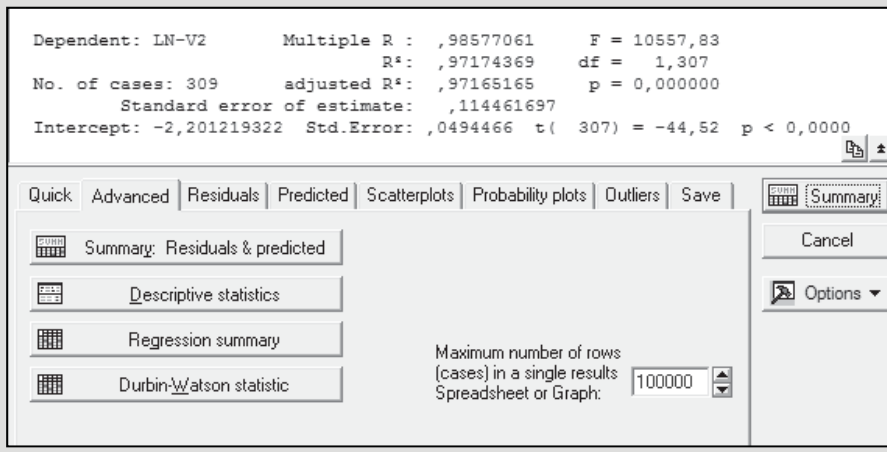


Se observa que existe homocedasticidad en los residuos, ya que no aumentan ni disminuyen en los extremos ni en el centro de la gráfica y, por tanto, son aproximadamente constantes a lo largo de todo el rango de datos.



Paso 19. Autocorrelación. Para determinar si no existe autocorrelación se utiliza el contraste de Durbin-Watson. En la pestaña «Advanced» se pulsa en «Durbin-Watson statistic».

CUADRO XI.1. *(Continuación)*[STATISTICA]



	Durbin-Watson d (Cuadro XI.1.sta) and serial correlation of residuals Include condition: v1="Molusco 3"	
	Durbin-Watson d	Serial Corr.
Estimate	1,769596	0,111394

El valor del estadístico de contraste es $\hat{d}$ = 1,769. Las tablas de Durbin-Watson (Tabla 7 del Apéndice I o archivo **Tabla 7.doc**) para n = 309, α = 0,05 y una variable independiente (k = 1) muestran $d_L \approx 1{,}81$ y $d_U \approx 1{,}82$.

- Si $0 < \hat{d} < d_L$ se rechaza H_0 y aceptamos la existencia de autocorrelación positiva.
- Si $d_L < \hat{d} < d_U$ el contraste no es concluyente.
- Si $d_U < \hat{d} < 4 - d_U$ se acepta H_0 y no hay autocorrelación.
- Si $4 - d_U < \hat{d} < 4 - d_L$ el contraste no es concluyente.
- Si $4 - d_L < \hat{d} < 4$ se rechaza H_0 y aceptamos la existencia de autocorrelación negativa.

Por lo tanto, como $\hat{d}$ (1,769) < d_L *(1,82) se rechaza* H_0 y existe autocorrelación positiva entre los residuos.

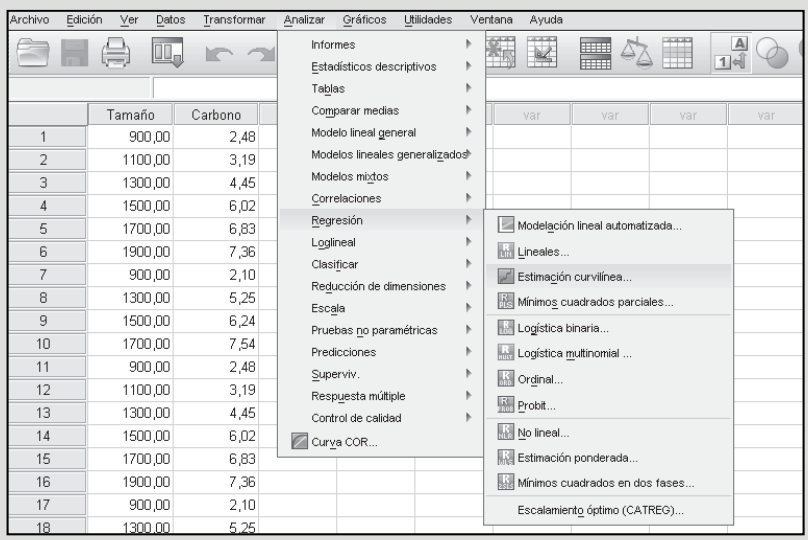
En cualquier caso, el hecho de que no se cumplan los 3 requisitos no invalida el modelo ya que es muy predictivo, el 97,2% de la variación observada en el peso se puede explicar en función de la longitud de la concha. La existencia de un pequeño grado de autocorrelación en los residuos indica que quizás debemos buscar alguna variable explicativa adicional para completar el modelo.

CUADRO XI.1. *(Continuación)*[SPSS]

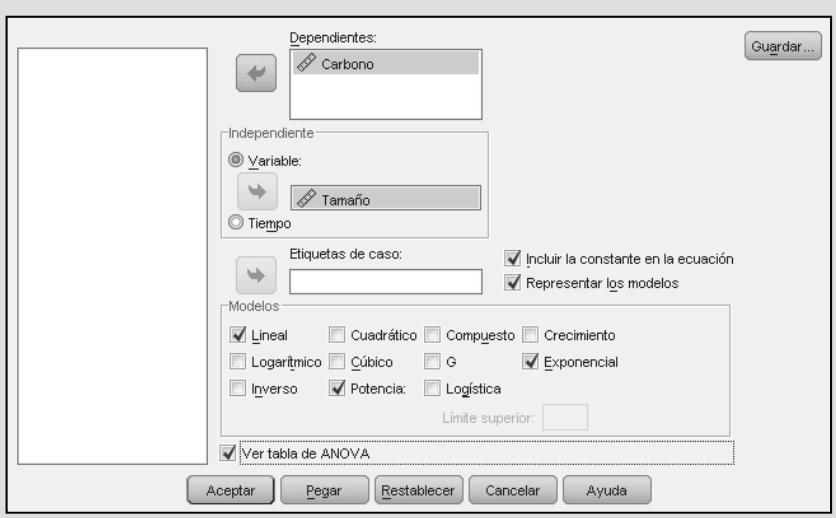
Aplicación con SPSS

EJEMPLO. En el archivo **Cuadro XI.1.sav** se muestra el tamaño de los utrículos (órganos encargados de capturar las presas) de una planta carnívora y su contenido en carbono (µg por utrículo). Se pretende determinar la ecuación que mejor se ajusta a los datos. Utilizaremos el programa SPSS.

Paso 1. Después de introducir nuestras variables, hay que entrar en la sección «Analizar» ⇨ «Regresión» ⇨ «Estimación curvilínea…».



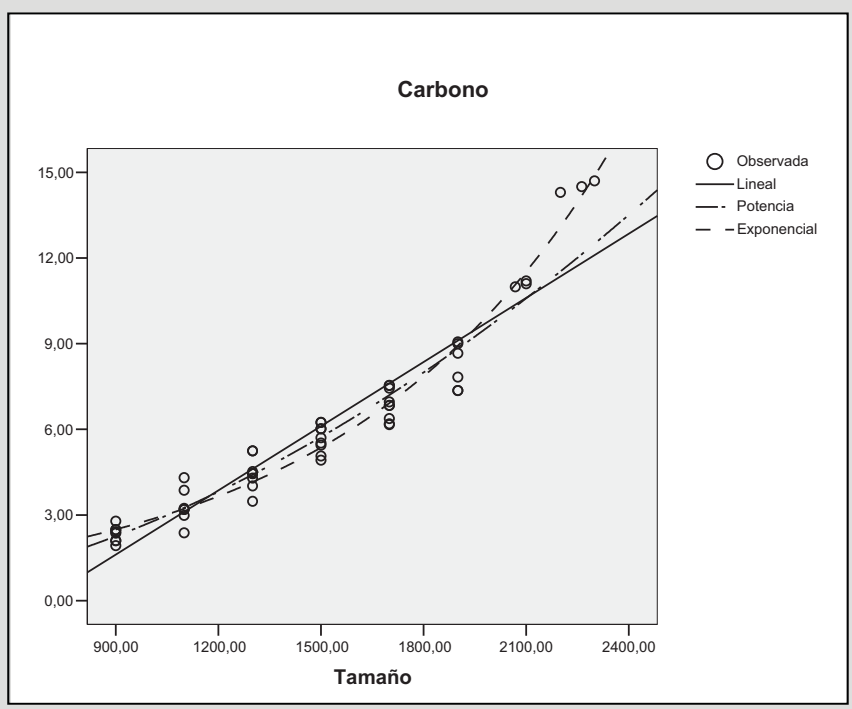
Paso 2. Nos saldrá la siguiente ventana en la que ponemos la variable dependiente «Carbono» y la independiente «Tamaño».



CUADRO XI.1. *(Continuación)*[SPSS]

Luego indicamos los tipos de funciones que queremos ver en el ajuste. En nuestro caso hemos seleccionado en «Modelos» ⇨ «Lineal», «Exponencial» y «Potencia». Es importante que siempre se incluya la constante de la ecuación (el punto de corte de la función con el eje *Y* cuando *X* es cero). Solo en los casos en que sepamos que cuando *X* es cero *Y* también tiene que ser cero, no incluiremos la constante de la ecuación. En nuestro caso lo incluimos y por ello marcamos «Incluir la constante en la ecuación». Al marcar «Representar los modelos» obtendremos la representación gráfica de los resultados. Por último, es importante marcar «Ver tabla de ANOVA» porque obtendremos información sobre el grado de significación del modelo.

Paso 3. Al «Aceptar» se obtiene el gráfico siguiente y los resultados que se muestran a continuación.



Para la ecuación Lineal:

Resumen del modelo

R	R cuadrado	R cuadrado corregida	Error típico de la estimación
,949	,901	,899	1,001

La variable independiente esTamaño.

CUADRO XI.1. *(Continuación)*[SPSS]

ANOVA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Regresión	472,192	1	472,192	471,241	,000
Residual	52,105	52	1,002		
Total	524,297	53			

La variable independiente esTamaño.

Coeficientes

	Coeficientes no estandarizados		Coeficientes estandarizados		
	B	Error típico	Beta	t	Sig.
Tamaño	,007	,000	,949	21,708	,000
(Constante)	-5,132	,528		-9,716	,000

Para la ecuación Potencial:

Resumen del modelo

R	R cuadrado	R cuadrado corregida	Error típico de la estimación
,973	,946	,945	,122

La variable independiente esTamaño.

ANOVA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Regresión	13,682	1	13,682	916,879	,000
Residual	,776	52	,015		
Total	14,458	53			

La variable independiente esTamaño.

Coeficientes

	Coeficientes no estandarizados		Coeficientes estandarizados		
	B	Error típico	Beta	t	Sig.
ln(Tamaño)	1,825	,060	,973	30,280	,000
(Constante)	9,12E-006	,000		2,283	,027

La variable dependiente es ln(Carbono).

CUADRO XI.1. *(Continuación)*[SPSS]

Y por último para la ecuación Exponencial:

Resumen del modelo

R	R cuadrado	R cuadrado corregida	Error típico de la estimación
,973	,947	,946	,121

La variable independiente esTamaño.

ANOVA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Regresión	13,691	1	13,691	927,896	,000
Residual	,767	52	,015		
Total	14,458	53			

La variable independiente esTamaño.

Coeficientes

	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
	B	Error típico	Beta		
Tamaño	,001	,000	,973	30,461	,000
(Constante)	,791	,051		15,602	,000

La variable dependiente es ln(Carbono).

Paso 4. En la interpretación de los resultados es importante tener en cuenta los siguientes criterios:

1. En la tabla de «Resumen del modelo», el «R» (coeficiente de correlación de Pearson) o «R cuadrado» nos indica el grado de correlación que existe entre las variables. El valor de r oscila entre -1 y 1, y el valor de r^2 oscila entre 0 y 1. Conforme el valor de r^2 sea más cercano a 1, mayor será la correlación y mejor será el ajuste. El r^2 es 0,9 para la lineal y 0,95 para la potencial y exponencial. Por tanto, el grado de ajuste más alto se obtiene usando las ecuaciones exponencial o potencial.

2. En la tabla «ANOVA» obtenemos otro resultado importante, el valor del estadístico F y el nivel de significación «Sig.». Como vemos en los tres modelos la probabilidad es menor de 0,001 y, por tanto, las regresiones son significativas. El valor de F es mayor en la exponencial (927,9) que en la potencial (916,8) y la lineal (471,2), siendo los grados de libertad en ambos casos los mismos (1, 52), indicando que la relación exponencial se ajusta mejor a nuestros datos.

CUADRO XI.1. *(Continuación)*[SPSS]

3. En la tabla de «Coeficientes» se muestran otros resultados importantes: las estimaciones de las constantes o parámetros que definen la ecuación:

Lineal: $Carbono = -5{,}132 + 0{,}00749 * Tamaño$

Potencial: $Carbono = 9{,}12 * 10^{-6} * Tamaño^{1{,}825}$

o también $\ln(Carbono) = -11{,}605 + 1{,}825 * \ln(Tamaño)$

Exponencial: $Carbono = 0{,}7906 * e^{0{,}00127 * Tamaño}$

o también $\ln(Carbono) = -0{,}2349 + 0{,}00127 * Tamaño$

4. El último resultado importante a tener en cuenta es el grado de significación asociado a las estimaciones de las constantes de la ecuación, que se muestra en la tabla de "Coeficientes". En el caso de la ecuación lineal y la exponencial, tanto la constante como la pendiente tienen una $p < 0{,}001$, indicando que son significativas. En el caso de la ecuación potencial, también es significativa la pendiente ($p < 0{,}001$) y la constante, pero en este caso con $p = 0{,}027$.

Paso 5. La conclusión es que tanto la ecuación Exponencial como la Potencial se ajustan mejor a nuestros datos. Se escoge la Potencial ya que normalmente las relaciones tamaño/biomasa son de ese tipo. La forma de expresar nuestros resultados sería la siguiente: La relación entre tamaño y carbono de los utrículos es significativa ($r^2 = 0{,}95$, $F_{1,52} = 916$, $p < 0{,}001$) y se ajusta a la ecuación Potencial que se mostró anteriormente.

Sin embargo, para que pueda ser utilizado el estadístico *F* y su nivel de significación, deben cumplirse los requisitos mencionados anteriormente en la sección XI.1.1 (homocedasticidad de los residuos, que éstos tengan una distribución Normal y que no exista autocorrelación entre ellos). Para comprobarlo es necesario realizar el modelo de regresión lineal.

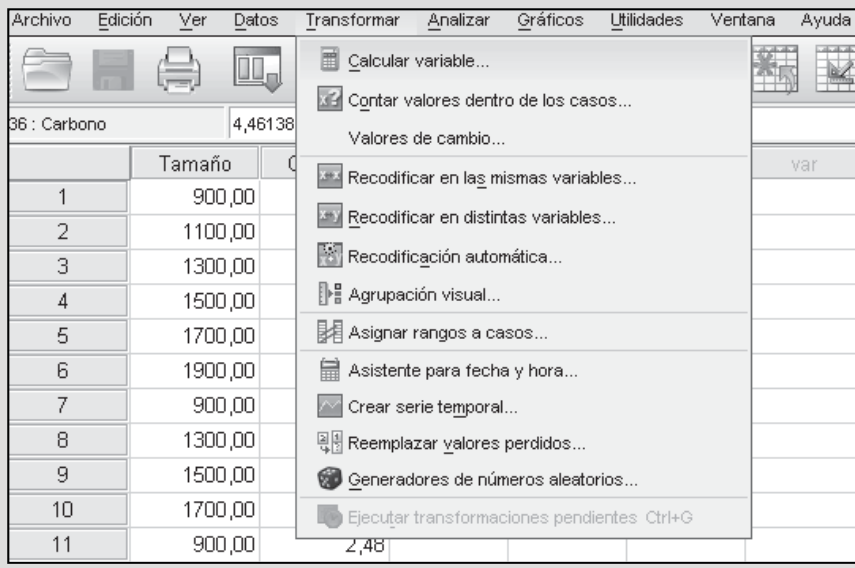
Paso 6. Como el modelo que mejor se ajusta es la ecuación Potencial, para poder aplicar el modelo de regresión lineal es necesario en primer lugar transformar nuestros datos para que la relación entre ambas variables sea lineal. En la Tabla XI.1 se muestran las transformaciones que hay que hacer para que la relación entre variables sea lineal.

Tabla XI.1. Transformaciones para conseguir relación lineal.

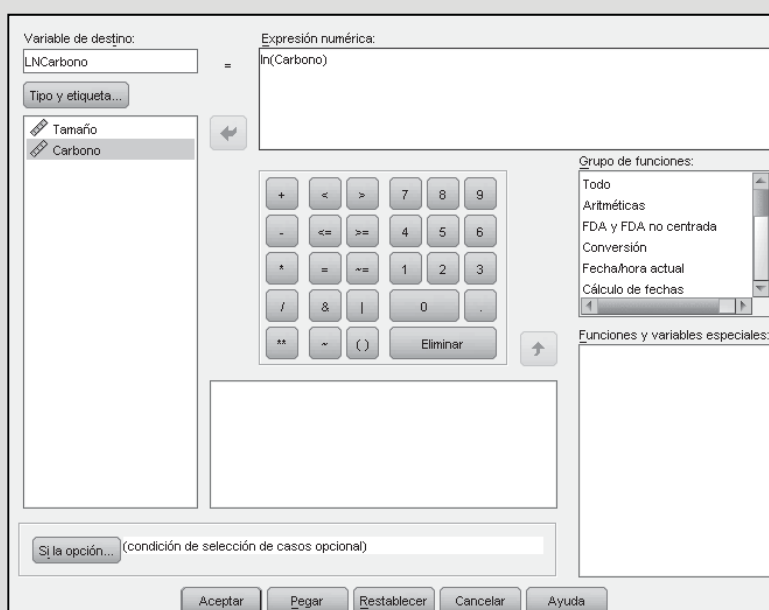
	Modelo	Transformación *x*	*Transformación y*
Logarítmica	$y = a + b\ln x$	$t(x) = \ln(x)$	$t(y) = y$
Potencial	$y = ax^b$	$t(x) = \ln(x)$	$t(y) = \ln(y)$
Exponencial	$y = ae^{bx}$	$t(x) = x$	$t(y) = \ln(y)$
Curva-S	$y = e^{a+\frac{b}{x}}$	$t(x) = \frac{1}{x}$	$t(y) = \ln(y)$
Inversa	$y = a + \frac{b}{x}$	$t(x) = \frac{1}{x}$	$t(y) = y$

CUADRO XI.1. *(Continuación)*[SPSS]

Paso 7. La transformación se realiza en «Transformar» ⇨ «Calcular variable…».

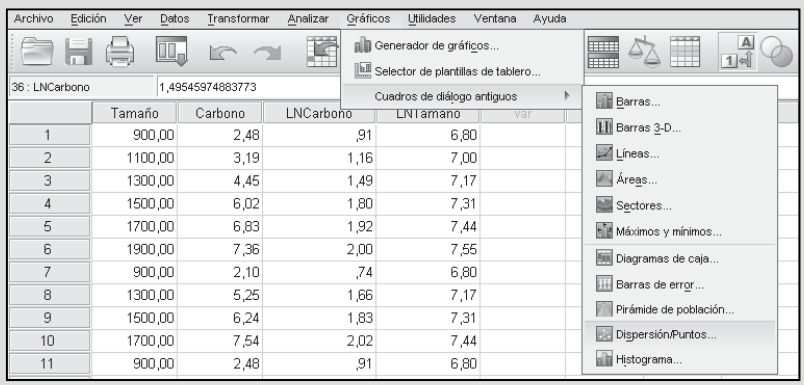


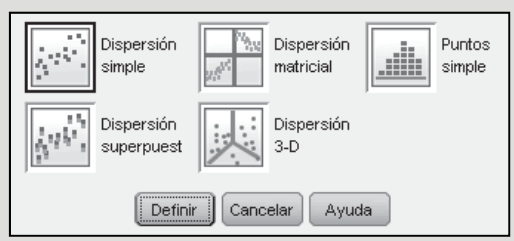
Paso 8. En la ventana que aparece, en «Variable de destino:» creamos una nueva variable «LNCarbono» y en «Expresión numérica:» ponemos la transformación que queremos realizar. Como el mejor ajuste lo obteníamos con la Potencial, hay que repetir el proceso y también calcular el logaritmo del «Tamaño».



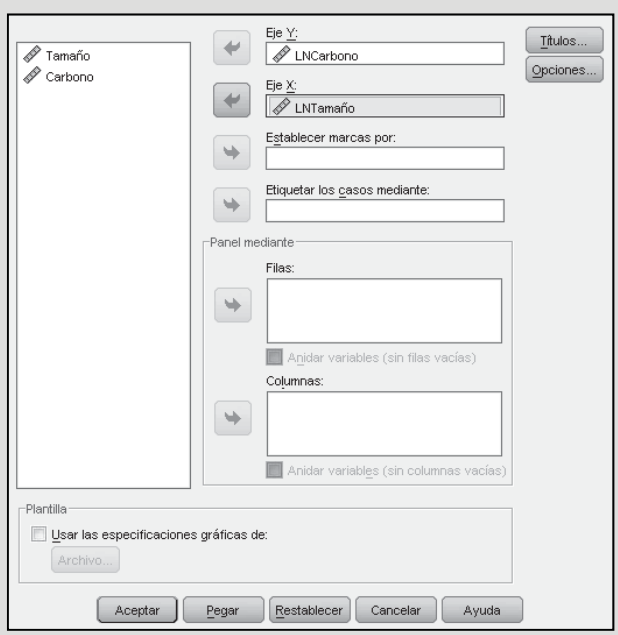
CUADRO XI.1. *(Continuación)*[SPSS]

Paso 9. Se representan los datos seleccionando «Gráficos» ⇨ «Dispersión/Puntos» y, en la siguiente ventana, se pulsa en «Dispersión simple».



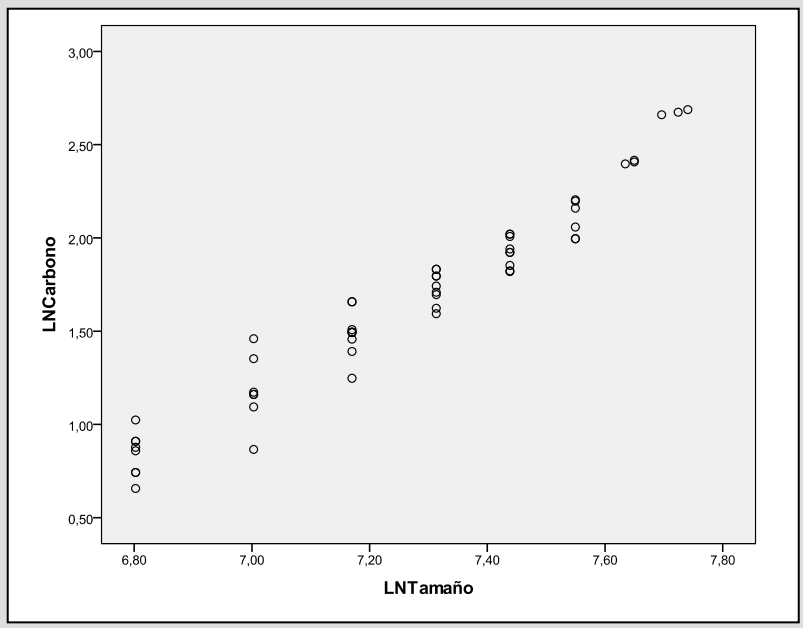


Por último, en el «Eje Y:» se introduce la variable transformada «LNCarbono» y en el «Eje X: » el «LNTamaño».

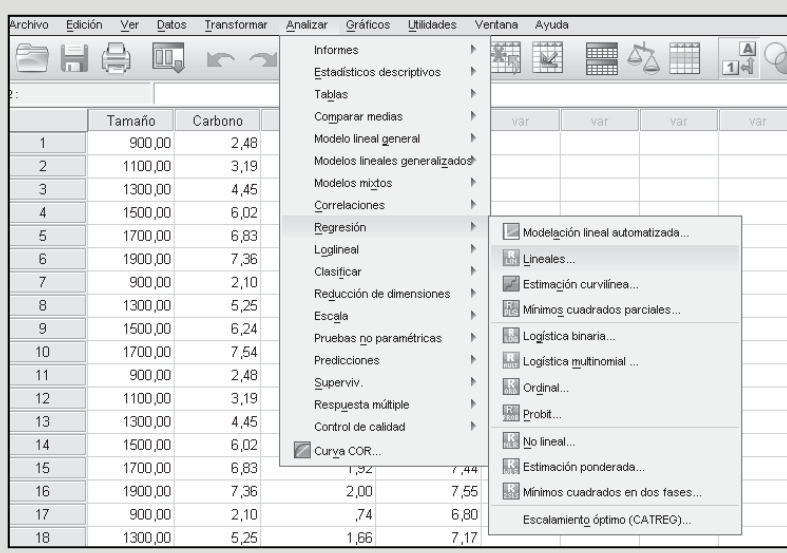


CUADRO XI.1. *(Continuación)*[SPSS]

Como se observa en la figura siguiente, al transformar los datos queda una relación lineal entre ambas variables.

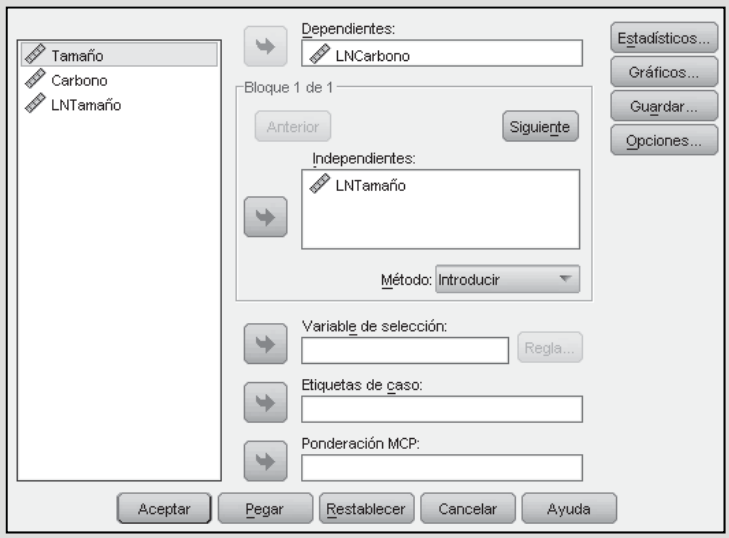


Paso 10. Para realizar la regresión lineal, dentro del menú principal nos vamos a «Analizar» ⇨ «Regresión» ⇨ «Lineales…».



Paso 11. En la ventana que aparece introducimos como variable dependiente «LNCarbono» y como independiente «LNTamaño».

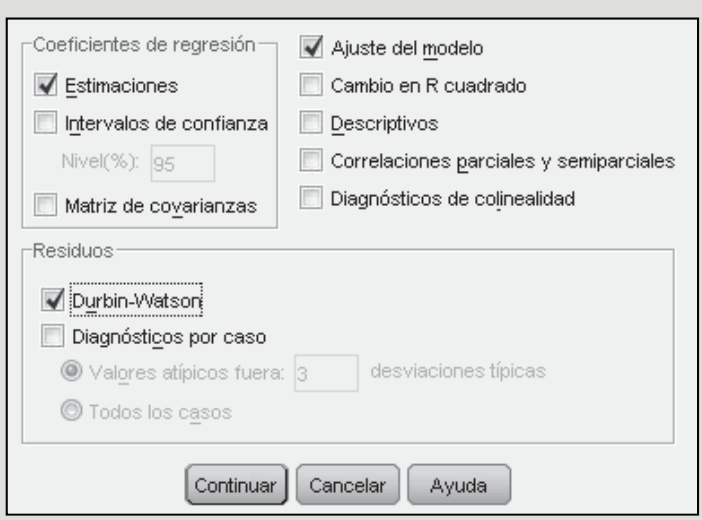
CUADRO XI.1. *(Continuación)*[SPSS]



En «Variable de selección:» se puede introducir una variable que permita seleccionar solo determinados casos para el análisis. Por ejemplo, poniendo 0 y 1 en la variable selección y diciendo que sólo se trabaje con los casos que tengan el código 1.

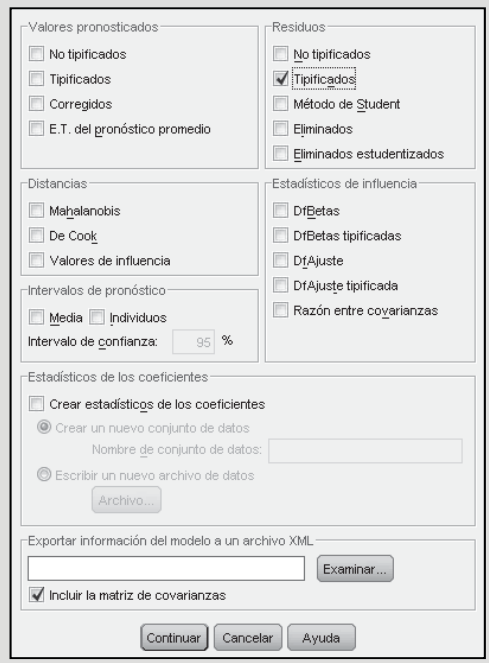
«Etiquetas de caso» permite diferenciar los casos a la hora de realizar una representación.

Paso 12. En la ventana del paso 11 entramos en «Estadísticos…» y en la ventana que aparece seleccionamos en «Residuos» el test «Durbin-Watson» que permite ver si existe autocorrelación entre los residuos. No tiene sentido ver la colinealidad y seleccionar «Diagnósticos de colinealidad» porque sólo hay una variable independiente. Dejamos seleccionado también «Estimaciones» y «Ajuste del modelo».

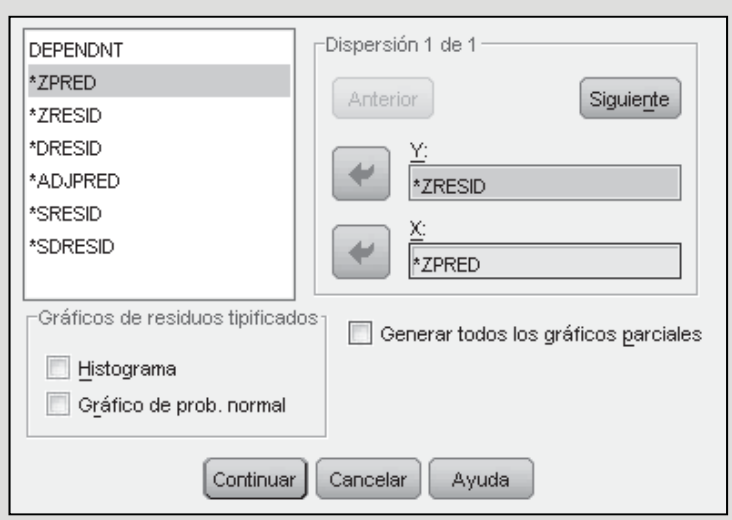


CUADRO XI.1. *(Continuación)*[SPSS]

Paso 13. En la ventana del paso 11, en el icono «Guardar…» aparece la siguiente ventana donde seleccionamos en «Residuos» los «Tipificados» que son los residuos que usaremos para estudiar la distribución Normal y homocedasticidad.

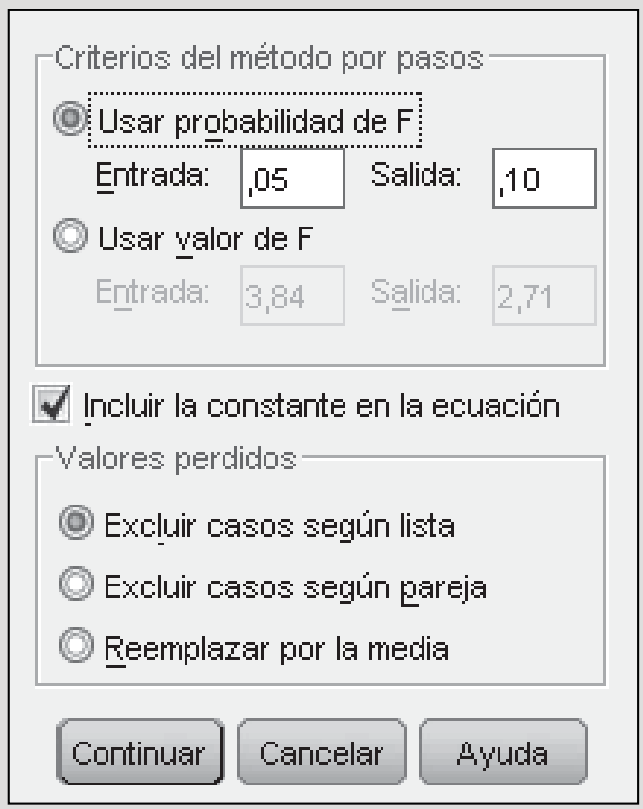


Paso 14. En la ventana del paso 11 si pulsamos en «Gráficos…» nos aparece la siguiente ventana donde podemos especificar que se realice el gráfico entre «ZRESID» (residuos tipificados) y «ZPRED» (valores pronosticados tipificados). Esta gráfica será necesaria para ver la homocedasticidad de los residuos.



CUADRO XI.1. *(Continuación)*[SPSS]

Paso 15. En la ventana del paso 11 al pulsar en «Opciones…» nos aparece la siguiente ventana donde podemos modificar el nivel de significación para que una variable entre o salga en la regresión por pasos, incluir o no incluir la intersección en la ecuación (lo que se denomina en el programa constante) y excluir casos según alguno de los criterios que se mencionan a continuación:



- *Excluir casos según lista*. Sólo se incluirán en el análisis los casos con valores válidos para todas las variables.

- *Excluir casos según pareja*. Los casos con datos completos para la pareja de variables correlacionadas se utilizan para calcular el coeficiente de correlación en el cual se basa el análisis de regresión. Los grados de libertad se basan en el N mínimo de las parejas.

- *Reemplazar por la media*. Se emplean todos los casos en los cálculos, sustituyendo las observaciones perdidas por la media de la variable.

CUADRO XI.1. *(Continuación)*[SPSS]

Paso 16. En la tabla «Resumen del modelo» y en la tabla de «ANOVA» se observa como el valor de r^2 = 0,95 y el valor de F = 916,9 son iguales a los que se obtenían con la ecuación Potencial (ver paso 4.3).

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,973[a]	,946	,945	,12216	,911

a. Variables predictoras: (Constante), LNTamaño
b. Variable dependiente: LNCarbono

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	13,682	1	13,682	916,879	,000[a]
	Residual	,776	52	,015		
	Total	14,458	53			

a. Variables predictoras: (Constante), LNTamaño
b. Variable dependiente: LNCarbono

En la tabla de «Coeficientes», que se muestra a continuación, se observa que, tanto la constante de la ecuación (-11,605) como la pendiente (1,825) son significativas con una $p < 0,001$.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	-11,605	,438		-26,489	,000
	LNTamaño	1,825	,060	,973	30,280	,000

a. Variable dependiente: LNCarbono

La forma de expresar nuestros resultados sería la siguiente: la relación entre tamaño y carbono de los utrículos es significativa (r^2 = 0,95, $F_{1,52}$ = 916,8, $p < 0,001$) y la ecuación que relaciona ambas variables es:

$$\ln(Carbono) = -11{,}605 + 1{,}825 * \ln(Tamaño)$$

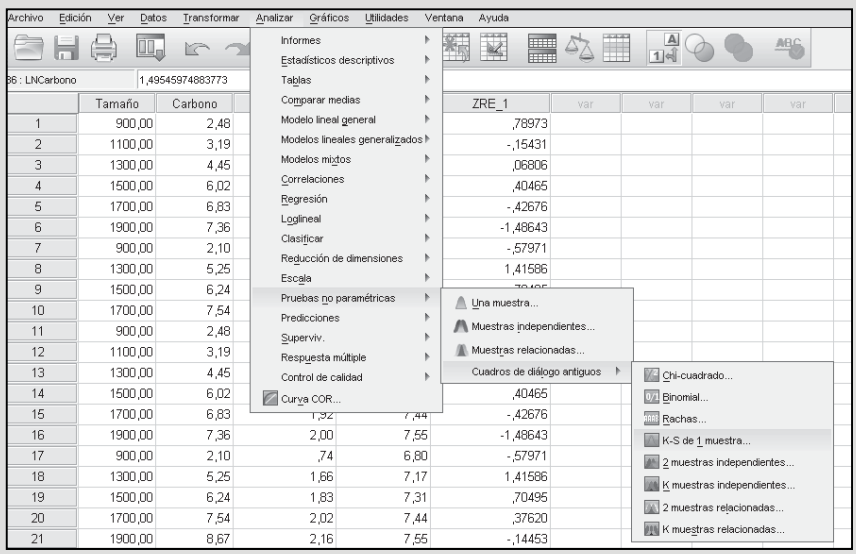
o en la otra forma de expresarla:

$$Carbono = 9{,}123 * 10^{-6} * Tamaño^{1{,}825}$$

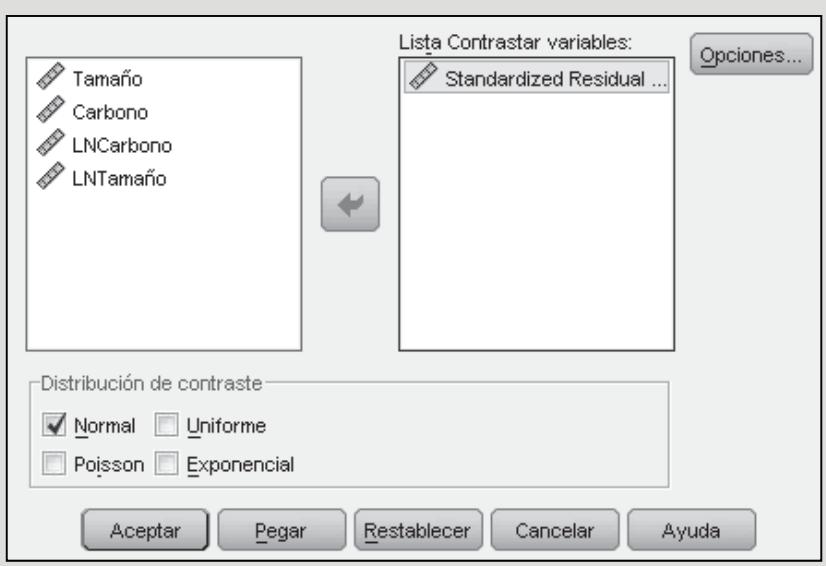
Ahora que ya hemos conseguido la linealidad entre las variables y hemos realizado la regresión lineal podemos ver si es correcto usar el estadístico F y su nivel de significación, examinando los requisitos de distribución Normal y homocedasticidad de los residuos y que no exista autocorrelación entre ellos.

CUADRO XI.1. *(Continuación)*[SPSS]

Paso 17. Distribución normal de los residuos. Volviendo a la página principal veremos que se ha creado una nueva variable que contiene los residuos tipificados «ZRE_1». Para ver la distribución Normal de los residuos entramos en «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «KS de 1 muestra», para hacer el test de Kolmogorov-Smirnov.



Nos aparece la siguiente ventana en la cual, dentro de «Lista Contrastar variables:», introducimos la variable de la que queremos ver si tiene una distribución Normal, en este caso «Standardized Residual…».



CUADRO XI.1. *(Continuación)*[SPSS]

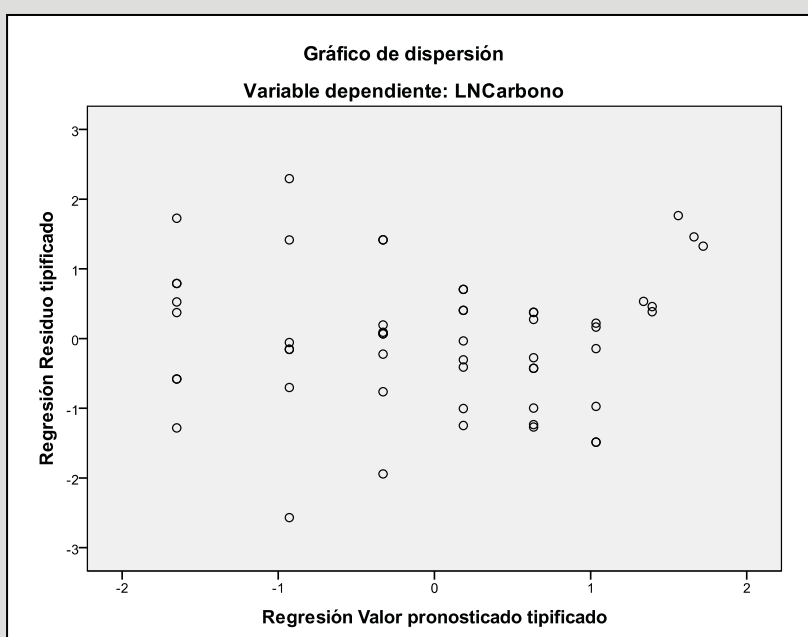
Al «Aceptar» nos aparece la siguiente tabla de resultados «Prueba de Kolmogorov-Smirnov para una muestra» en la que se observa que p = 0,935 y, por lo tanto, se cumple la hipótesis nula de que los residuos se ajustan a una distribución normal.

Prueba de Kolmogorov-Smirnov para una muestra

		Standardized Residual
N		54
Parámetros normales[a,b]	Media	,0000000
	Desviación típica	,99052111
Diferencias más extremas	Absoluta	,073
	Positiva	,073
	Negativa	-,058
Z de Kolmogorov-Smirnov		,537
Sig. asintót. (bilateral)		,935

a. La distribución de contraste es la Normal.
b. Se han calculado a partir de los datos.

Paso 18. Homocedasticidad de los residuos. Uno de los resultados obtenidos es el gráfico que se muestra a continuación (y que habíamos especificado en el paso 14) entre los residuos tipificados y las predicciones. Para que exista estabilidad en la variabilidad de los residuos, ésta debe ser más o menos constante, no aumentando ni disminuyendo en los extremos o el centro de la gráfica. Vemos como en este caso los residuos varían de forma similar a lo largo de todo el recorrido y, por tanto, aceptamos que existe homocedasticidad en los residuos.



CUADRO XI.1. *(Continuación)*^{SPSS}

Paso 19. Autocorrelación de los residuos. Para determinar si no existe autocorrelación en los residuos se utiliza el contraste de Durbin-Watson. En el paso 16, en la tabla «Resumen del modelo», se ve el valor del contraste $\hat{d} = 0{,}911$. Las tablas de Durbin-Watson (Tabla 7 del Apéndice I o archivo **Tabla 7.doc**) para $n = 54$ y $\alpha = 0{,}05$ proporcionan los valores aproximados de $d_L = 1{,}527$ y $d_U = 1{,}601$.

- Si $0 < \hat{d} < d_L$ se rechaza H_0 y aceptamos la existencia de autocorrelación positiva.
- Si $d_L < \hat{d} < d_U$ el contraste no es concluyente.
- Si $d_U < \hat{d} < 4 - d_U$ se acepta H_0 y no hay autocorrelación.
- Si $4 - d_U < \hat{d} < 4 - d_L$ el contraste no es concluyente.
- Si $4 - d_L < \hat{d} < 4$ se rechaza H_0 y aceptamos la existencia de autocorrelación negativa.

Por lo tanto, como $\hat{d}(0{,}911) < d_L$ *(1,527) se rechaza* H_0 y existe autocorrelación positiva entre los residuos. La autocorrelación aparece generalmente porque las medidas se toman a lo largo del tiempo y los errores se deben a los procedimientos utilizados. Una posible solución consiste en aleatorizar el proceso de medida en la fase de toma de datos.

En conclusión, aunque el modelo de regresión es muy significativo, no se debe utilizar el estadístico *F* ni el grado de significación que se obtiene, porque existe autocorrelación (los residuos no son independientes).

XI.1.3. REGRESIÓN MÚLTIPLE

Hasta ahora hemos visto regresiones en las que había una sola variable independiente. Sin embargo, es muy frecuente que necesitemos estudiar si nuestra variable dependiente está relacionada con más de una variable. En este caso, es necesario usar la regresión múltiple.

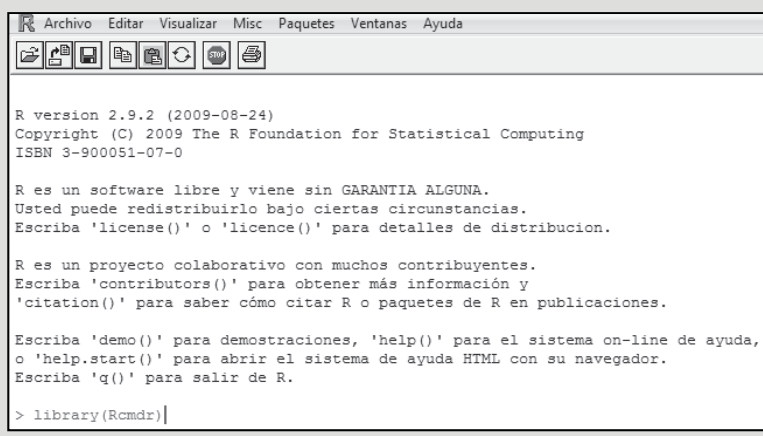
Al igual que ocurría con la regresión simple, las relaciones entre variables pueden ser lineales o no lineales. En este apartado vamos a tratar sólo con aquellas situaciones en las que, aunque algunas variables no tengan relación lineal, es posible obtener la linealidad mediante alguna transformación. En el siguiente apartado se explicará la forma de obtener la regresión cuando la relación no es lineal.

CUADRO XI.2. Regresión múltiple

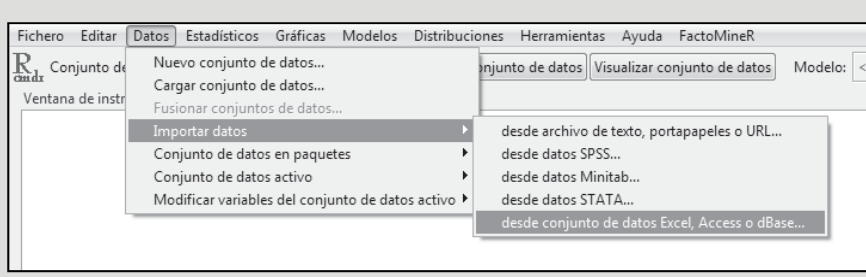
Aplicación con R.

EJEMPLO. En el archivo **Cuadro XI.2.xls** hay datos promedios a lo largo de varios años, para diferentes latitudes y longitudes en zonas de afloramiento, de clorofila *a* (mgm^{-3}), porcentaje que ocupa la plataforma continental, temperatura (ºC), transporte de Ekman ($m^3km^{-1}s^{-1}$, indicador de afloramiento) y turbulencia (m^3s^{-3}). Los datos incluyen las transformaciones logarítmicas de las variables. El objetivo es determinar si es posible explicar la variación observada en la clorofila *a* en función de las otras variables. El análisis se realizará solo para las zonas de afloramiento del hemisferio Norte: Canarias y California.

Paso 1. Son necesarios los paquetes "MASS", "corpcor", "lmtest", "nortest", "stepwise", "leaps", "boot", "mitools" y "Rcmdr". También es necesario el paquete "relaimpo", que para su instalación es mejor descargar la carpeta del la página web: http://prof.beuth-hochschule.de/groemping/software/relaimpo-relative-importance-of-regressors/download-package-non-us-version/, y guardarla en la carpeta «library» de R. Iniciamos R y cargamos el paquete «Rcmdr» introduciendo la orden «library(Rcmdr)».

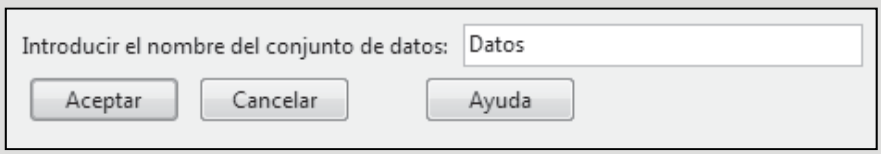

Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> library(Rcmdr)

Paso 2. Después de haber seleccionado el directorio de trabajo, como se explica en el Apéndice II, vamos a «Datos» ⇨ «Importar datos» ⇨ «desde conjunto de datos Excel, Access o dBase».


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda FactoMineR
Nuevo conjunto de datos...
Cargar conjunto de datos...
Fusionar conjuntos de datos...
Importar datos
Conjunto de datos en paquetes
Conjunto de datos activo
Modificar variables del conjunto de datos activo
desde archivo de texto, portapapeles o URL...
desde datos SPSS...
desde datos Minitab...
desde datos STATA...
desde conjunto de datos Excel, Access o dBase...
Visualizar conjunto de datos
Modelo:
Ventana de instr

CUADRO XI.2. *(Continuación)*[R]

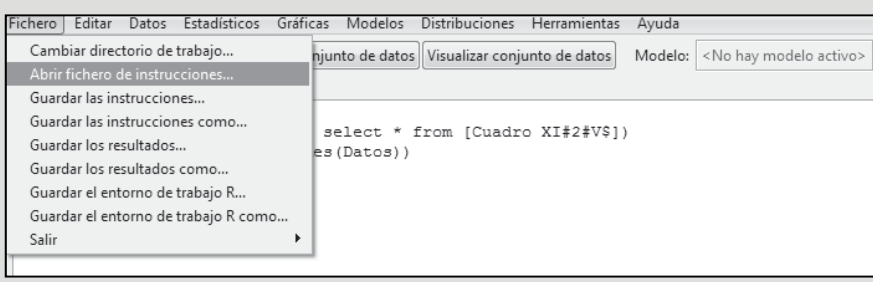
Paso 3. Aparece un pequeño cuadro de diálogo para introducir el nombre del conjunto de datos, en el que dejamos el nombre por defecto «Datos» y pulsamos «Aceptar». Posteriormente aparecen los archivos del directorio de trabajo, en el que elegimos **Cuadro XI.2.xls** y pulsamos «Abrir».



Paso 4. En R Commander pulsamos en «Visualizar conjunto de datos» lo que nos permite inspeccionar nuestros datos.

	Región	Lat.N	Long.E	Clorofila.a	Plataforma.continental	Ekman	Temperatura	Turbulencia	LnClorofila.a
1	Canarias	35.5	-6.5	0.24	30.83	324.9	19.1	646.9	-1.427
2	Canarias	34.5	-7.5	0.15	2.78	404.8	20.0	421.2	-1.897
3	Canarias	33.5	-8.5	0.31	60.00	524.8	20.0	488.3	-1.171
4	Canarias	32.5	-9.5	0.51	59.32	762.9	19.6	641.9	-0.673
5	Canarias	31.5	-10.5	0.37	7.64	1214.3	19.5	839.4	-0.994
6	Canarias	30.5	-10.5	0.52	0.69	1227.2	18.7	710.3	-0.654
7	Canarias	29.5	-11.5	0.25	7.64	1158.1	19.7	563.8	-1.386
8	Canarias	28.5	-12.5	0.64	60.99	1287.4	19.7	638.3	-0.446
9	Canarias	27.5	-13.5	0.89	37.93	1478.4	20.1	682.6	-0.117
10	Canarias	26.5	-14.5	2.05	21.14	1509.6	20.3	763.7	0.718
11	Canarias	25.5	-15.5	2.50	59.72	1699.0	20.3	756.3	0.916

Paso 5. A continuación se carga el archivo de instrucciones **Cuadro XI.2.R** seleccionando «Fichero» ⇨ «Abrir fichero de instrucciones…» y no se guarda la salida.



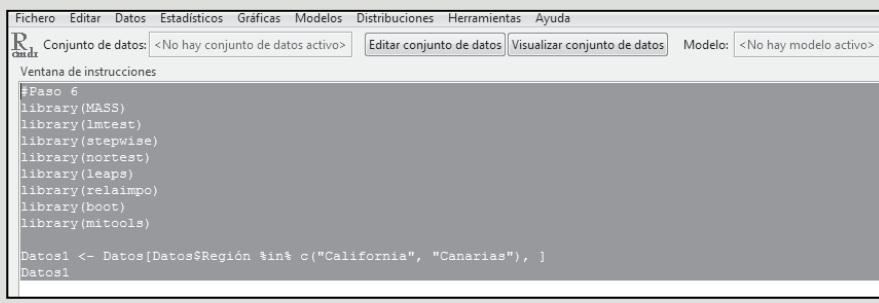
Paso 6. Las primeras instrucciones cargan los paquetes necesarios y las siguientes seleccionan, del archivo de datos, sólo las regiones con las que se quiere trabajar (Canarias y California).

```
library(MASS)
library(lmtest)
library(stepwise)
library(nortest)
library(leaps)
library(relaimpo)
library(boot)
library(mitools)
```

CUADRO XI.2. *(Continuación)*[R]

Datos1 <- Datos[Datos$Región %in% c("California", "Canarias"),] *# selecciona en el archivo «Datos», en la variable Región», sólo las zonas de «California» y «Canarias», y crea una nueva base de datos que se llama «Datos1»*
Datos1 *# muestra los datos seleccionados en la ventana de resultados*

Se seleccionan todas las instrucciones del paso 6 y se pulsa en «Ejecutar». Así será el proceso con el resto de pasos, es decir, se seleccionan las instrucciones de cada paso y se pulsa en «Ejecutar».



En la ventana aparece la nueva tabla de datos con la que se va a trabajar.

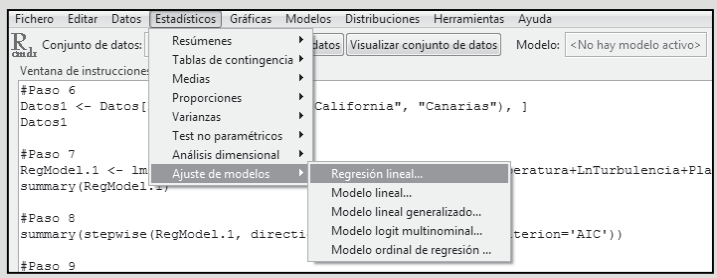
```
> Datos1
       Región Lat.N Long.E Clorofila.a Plataforma.continental  Ekman Temperatura Turbulencia
1    Canarias  35.5   -6.5        0.24                  30.83  324.9        19.1       646.9
2    Canarias  34.5   -7.5        0.15                   2.78  404.8        20.0       421.2
3    Canarias  33.5   -8.5        0.31                  60.00  524.8        20.0       488.3
4    Canarias  32.5   -9.5        0.51                  59.32  762.9        19.6       641.9
5    Canarias  31.5  -10.5        0.37                   7.64 1214.3        19.5       839.4
6    Canarias  30.5  -10.5        0.52                   0.69 1227.2        18.7       710.3
7    Canarias  29.5  -11.5        0.25                   7.64 1158.1        19.7       563.8
8    Canarias  28.5  -12.5        0.64                  60.99 1287.4        19.7       638.3
9    Canarias  27.5  -13.5        0.89                  37.93 1478.4        20.1       682.6
10   Canarias  26.5  -14.5        2.05                  21.14 1509.6        20.3       763.7
11   Canarias  25.5  -15.5        2.50                  59.72 1699.0        20.3       756.3
12   Canarias  24.5  -15.5        2.68                 100.00 1561.7        19.2       644.1
13   Canarias  23.5  -16.5        2.55                  87.05 1647.8        19.6       663.7
14   Canarias  22.5  -16.5        4.43                 100.00 1979.4        19.1       772.1
15   Canarias  21.5  -17.5        3.15                  39.16 2297.7        20.0       890.7
47 California  39.5 -124.5        0.83                   0.00 1224.3        11.9      1643.9
48 California  38.5 -123.5        1.65                  29.79 1236.9        11.6      1295.2
49 California  37.5 -123.5        1.17                   9.72 1044.5        13.2       999.4
50 California  36.5 -122.5        1.23                   2.13 1082.4        13.1       902.1
51 California  35.5 -121.5        1.18                   2.59 1307.2        13.5       961.2
52 California  34.5 -120.5        1.96                  34.94 1352.2        13.7       973.2
53 California  33.5 -119.5        0.80                   4.48  866.4        16.7       507.7
54 California  33.5 -118.5        0.62                  12.61  319.9        16.9       188.5
55 California  32.5 -117.5        0.44                   2.56  409.2        17.6       215.2
```

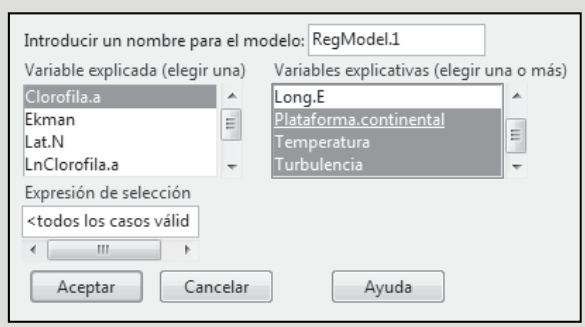
Paso 7. A continuación se realiza la regresión múltiple. Se selecciona la variable dependiente, en este caso la «Clorofila.a», y las variables independientes, que serían el resto de variables. En primer lugar realizaremos el análisis con la variable dependiente no transformada logarítmicamente, y como independientes se incluyen todas las variables, tanto transformadas como no transformadas, lógicamente a excepción de la transformación logarítmica de la clorofila y de las variables que informan sobre la latitud y longitud.

CUADRO XI.2. *(Continuación)*[R]

RegModel.1 <- lm(Clorofila.a~Ekman+LnEkman+LnPlataforma+LnTemperatura+LnTurbulencia+Plataforma.continental+Temperatura+Turbulencia, data=Datos1) *# realiza la regresión múltiple con la variable dependiente «Clorofila.a» y como independientes el resto de variables*
summary(RegModel.1)

También es posible realizar la regresión múltiple seleccionando «Estadísticos» ⇨ «Ajuste de modelos» ⇨ «Regresión lineal…», y en la ventana que aparece seleccionar la variable dependiente y las variables independientes.





Los resultados muestran que sólo algunas variables son significativas («Ekman», «LNEkman» y «Plataforma.continental»). Por tanto, es conveniente realizar una regresión por pasos.

```
lm(formula = Clorofila.a ~ Ekman + LnEkman + LnPlataforma + LnTemperatura +
    LnTurbulencia + Plataforma.continental + Temperatura + Turbulencia,
    data = Datos1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67990 -0.32587 -0.01719  0.15606  1.29063

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            15.6364331 29.4055361   0.532 0.600479
Ekman                   0.0042051  0.0009173   4.584 0.000161 ***
LnEkman                -2.3689220  0.8392218  -2.823 0.010195 *
LnPlataforma           -0.1656337  0.1769331  -0.936 0.359844
LnTemperatura          -1.1220450 16.3596257  -0.069 0.945968
LnTurbulencia           0.6558416  0.9664395   0.679 0.504791
Plataforma.continental  0.0161028  0.0073279   2.197 0.039345 *
Temperatura            -0.1291476  0.9699081  -0.133 0.895339
Turbulencia            -0.0022158  0.0019405  -1.142 0.266374
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5154 on 21 degrees of freedom
Multiple R-squared: 0.8105, Adjusted R-squared: 0.7383
F-statistic: 11.23 on 8 and 21 DF,  p-value: 4.764e-06
```

CUADRO XI.2. *(Continuación)*[R]

Paso 8. La selección de variables paso a paso consiste en elegir las variables explicativas que deben ser incluidas en la regresión mediante algún criterio estadístico que asegure que solamente sean utilizadas las variables que mejoran la capacidad explicativa del modelo en su conjunto.

Existen básicamente dos procedimientos paso a paso: adelante y atrás. Con el primero –dirección adelante– el proceso se inicia sin ninguna variable (solo la constante) y se busca la variable independiente con mayor capacidad de explicación de la variabilidad de la variable dependiente. A continuación se busca, entre las restantes variables explicativas, aquella que combinada con la primera aumenta en mayor medida la capacidad del modelo, que es añadida a la regresión, y se continúa de este modo añadiendo variables sucesivamente, y mejorando progresivamente la capacidad explicativa del modelo, hasta que no se encuentra ninguna variable que la incremente significativamente, con lo que el proceso se detiene. Con el segundo –dirección atrás– el proceso empieza con todas las variables, y se van descartando sucesivamente aquellas que no aportan al modelo una capacidad significativa, hasta que solo permanecen las variables de interés. A menudo los resultados finales de ambos procedimientos son similares.

En cuanto al criterio utilizado para decidir cuando una variable aporta o no al modelo una capacidad explicativa significativa que justifique su introducción, se suele utilizar –además de la significación estadística de la variable considerada– una penalización por la complejidad del modelo, ya que en general es preferible un modelo sencillo con pocas variables a un modelo con muchas variables que solo incrementa ligeramente la capacidad explicativa. Esto ha dado lugar a dos criterios estadísticos mayoritariamente utilizados: Akaike's Information Criterion (AIC) y Bayesian Information Criterion (BIC); ambos pueden ser utilizados indistintamente.

La siguiente instrucción realiza el modelo de regresión por pasos.

summary(stepwise(RegModel.1, direction='forward/backward', criterion = 'AIC'))

```
Call:
lm(formula = Clorofila.a ~ Ekman + LnEkman + Plataforma.continental +
    Temperatura + Turbulencia, data = Datos1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72401 -0.33423  0.02127  0.15303  1.33085

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)            14.5762112  4.2286238   3.447  0.00210 **
Ekman                   0.0039112  0.0007812   5.006 4.09e-05 ***
LnEkman                -2.0573648  0.6867825  -2.996  0.00627 **
Plataforma.continental  0.0104845  0.0038246   2.741  0.01137 *
Temperatura            -0.1756710  0.0530657  -3.310  0.00294 **
Turbulencia            -0.0010895  0.0005087  -2.142  0.04258 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4947 on 24 degrees of freedom
Multiple R-squared: 0.8005, Adjusted R-squared: 0.7589
F-statistic: 19.26 on 5 and 24 DF,  p-value: 1.065e-07
```

CUADRO XI.2. *(Continuación)*[R]

Los resultados muestran que la regresión incluye las variables «Ekman», «LNEkman», «Plataforma.continental», «Temperatura» y «Turbulencia». Al existir una clara asociación entre «Ekman» y «LNEkman», es necesario eliminar una de las dos variables y, por ello, nos quedamos con «Ekman» ya que su probabilidad es menor (< 0,001) que «LNEkman» (0,006).

Paso 9. Se realiza de nuevo la regresión pero sólo incluyendo las variables «Ekman», «Plataforma.continental», «Temperatura» y «Turbulencia».

```
RegModel.2 <- lm(Clorofila.a~Ekman+Plataforma.continental+
Temperatura+Turbulencia, data=Datos1)
summary(RegModel.2)
```

```
Call:
lm(formula = Clorofila.a ~ Ekman + Plataforma.continental + Temperatura +
    Turbulencia, data = Datos1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.978346 -0.315582  0.005036  0.287775  1.226895

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            2.3155840  1.2209603   1.897   0.0695 .
Ekman                  0.0017177  0.0003128   5.492 1.05e-05 ***
Plataforma.continental 0.0118231  0.0043623   2.710   0.0120 *
Temperatura           -0.1471070  0.0599519  -2.454   0.0215 *
Turbulencia           -0.0011463  0.0005838  -1.963   0.0608 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5681 on 25 degrees of freedom
Multiple R-squared: 0.7259, Adjusted R-squared: 0.682
F-statistic: 16.55 on 4 and 25 DF,  p-value: 9.506e-07
```

Se observa que el modelo explica el 72,59% de la varianza observada en la concentración de clorofila.

Paso 10. A continuación es necesario repetir el proceso, pero utilizando la variable dependiente transformada «LnClorofila.a» y de nuevo como independientes el resto de variables.

```
RegModel.3 <- lm(LnClorofila.a~Ekman+LnEkman+LnPlataforma+Ln
Temperatura+LnTurbulencia+Plataforma.continental+Temperatura+Turbu
lencia, data=Datos1) # realiza la regresión múltiple incluyendo todas las
variables

summary(stepwise(RegModel.3, direction='forward/backward', criterion
='AIC')) # realiza la regression multiple por pasos
```

CUADRO XI.2. *(Continuación)*[R]

```
Call:
lm(formula = LnClorofila.a ~ Ekman + Temperatura + Turbulencia +
    Plataforma.continental, data = Datos1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89178 -0.20724  0.01083  0.10880  0.91956

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.4329725  0.8250910   2.949  0.00683 **
Ekman                   0.0016593  0.0002114   7.851 3.31e-08 ***
Temperatura            -0.2097657  0.0405138  -5.178 2.36e-05 ***
Turbulencia            -0.0013830  0.0003945  -3.505  0.00174 **
Plataforma.continental  0.0082759  0.0029479   2.807  0.00954 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3839 on 25 degrees of freedom
Multiple R-squared: 0.8194, Adjusted R-squared: 0.7906
F-statistic: 28.37 on 4 and 25 DF,  p-value: 5.735e-09
```

En el nuevo modelo el r^2 = 0,819, superior al anterior que era 0,725. Por tanto, nos quedamos con este modelo de regresión en el que la variable dependiente «LnClorofila.a» (*LnClo*) está asociada significativamente ($F_{4,25}$ = 28,37, p < 0,001) y de forma positiva con las variables «Ekman» (*E*) y «Plataforma.continental» (*P*), mientras que con una relación significativa negativa con «Temperatura» (*Te*) y «Turbulencia» (*Tu*), por medio de la siguiente ecuación:

$$LnCl = 2{,}43 + 0{,}00166*E - 0{,}2097*Te - 0{,}00138*Tu + 0{,}0082*P$$

Paso 11. Una vez elegido el mejor modelo realizamos por última vez la regresión múltiple, con las variables seleccionadas en el paso anterior, para poder tener los resultados en «Modelo.final».

```
Modelo.final<- lm(LnClorofila.a~Ekman+Plataforma.continental+
Temperatura+Turbulencia, data=Datos1)
```

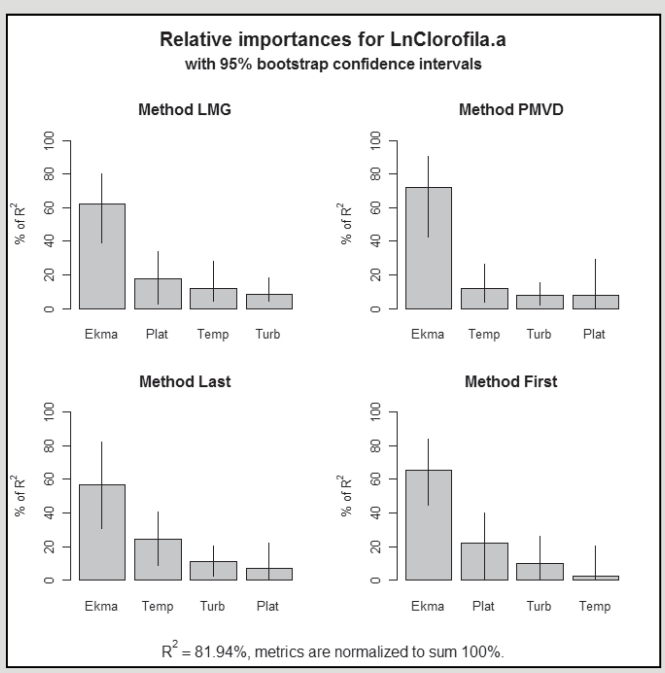
Paso 12. Las instrucciones que vienen a continuación permiten estimar, por cuatro métodos diferentes, la contribución de cada una de las variables independientes (el proceso de cálculo puede tardar varios segundos). Para que funcione el método «pmvd» es necesario descargar la carpeta "relaimpo" de la página web: http://prof.beuth-hochschule.de/groemping/software/relaimpo-relative-importance-of-regressors/download-package-non-us-version/, y guardarla en la carpeta «library» de R. Si da errores, simplemente se elimina «pmvd» de las instrucciones.

```
imprel<-calc.relimp(Modelo.final,type=c("lmg","pmvd","last","first"),rela
=TRUE)
boot <- boot.relimp(Modelo.final, b = 1000, type = c("lmg", "pmvd",
"last", "first"), rank = TRUE, diff = TRUE, rela = TRUE)
booteval.relimp(boot)
x11()
plot(booteval.relimp(boot,sort=TRUE)
```

CUADRO XI.2. *(Continuación)*[R]

Existen varios procedimientos para descomponer la varianza explicada en partes atribuidas a las diferentes variables explicativas, y se suele considerar que los más adecuados son «lmg» o «pmvd». En «lmg» la contribución de cada variable es distinta según el orden en que se consideran las variables en el modelo, y el método «lmg» calcula el promedio de contribuciones empleando todas las ordenaciones posibles. El método «pmvd» es una variante de «lmg» que calcula un promedio ponderado de las contribuciones de cada regresor a la varianza explicada (al r^2) considerando todas las ordenaciones, siendo las ponderaciones obtenidas de los datos (la diferencia con «lmg» está en la ponderación). Se considera el único procedimiento que garantiza cuatro propiedades deseables en la descomposición: que la suma de las contribuciones sea la varianza explicada total, que las contribuciones no sean negativas, que una variable con coeficiente de regresión nulo tenga contribución nula, y que si no es nulo no lo sea tampoco la contribución. Otros procedimientos calculan la aportación cuando la variable se incorpora al modelo en primer lugar «first», o en último lugar «last».

En todos los métodos se observa que el porcentaje mayor de varianza explicada se debe a la variable «Ekman» (más del 60% del r^2 en casi todos los métodos) y luego aparece la «Temperatura» o la «Plataforma.continental» dependiendo del método.



Paso 13. Las siguientes instrucciones calculan los residuos y los guardan en el archivo **Salida Cuadro XI.2.csv**. Los residuos son útiles para poder encontrar posibles *outliers* y para comprobar algunos de los requisitos de la regresión.

CUADRO XI.2. *(Continuación)*[R]

Residuos<-residuals(Modelo.final) *# calcula los residuos*
write.csv2(Residuos,file="Salida Cuadro XI.2.csv") *# graba los residuos en un archivo*

	x
1	0,2469332
2	-0,24685758
3	-0,10072842

Paso 14. Normalidad de los residuos. Las siguientes instrucciones permiten aplicar el test de Shapiro-Wilk y el de Kolmogorov-Smirnov con la corrección de Lilliefors a los residuos, para ver si tienen distribución Normal.

Shapiro<-shapiro.test(Residuos)
Shapiro
Lillie<-lillie.test(Residuos)
Lillie

Los residuos tienen una distribución Normal, ya que tanto en el test de Shapiro-Wilk como en el de Kolmogorov-Smirnov $p > 0,05$.

```
> Shapiro

        Shapiro-Wilk normality test

data:  Residuos
W = 0.9307, p-value = 0.05115


> Lillie<-lillie.test(Residuos)

> Lillie

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  Residuos
D = 0.1511, p-value = 0.07866
```

Paso 15. Autocorrelación. La comprobación de si existe autocorrelación se hace por medio de las siguientes instrucciones, que calculan y muestran, respectivamente, el valor del test de Durbin-Watson.

DW<-durbin.watson(Modelo.final)
DW

El valor de $p = 0,108$ indica que no existe autocorrelación, ya que $p > 0,05$. La presencia de autocorrelación en los residuos es un problema, y suele indicar que existe alguna variable no incluida en el modelo que debiera ser considerada.

```
> DW
 lag Autocorrelation D-W Statistic p-value
   1       0.1782888      1.623541   0.108
 Alternative hypothesis: rho != 0
```

CUADRO XI.2. *(Continuación)*[R]

Paso 16. Homocedasticidad. También se cumple el requisito de homocedasticidad de los residuos, ya que la probabilidad del test Breusch-Pagan es mayor que 0,05.

```
BreuschPagan<-bptest(Modelo.final)
BreuschPagan
```

```
> BreuschPagan

        studentized Breusch-Pagan test

data:  Modelo.final
BP = 7.3656, df = 4, p-value = 0.1178
```

Paso 17. Colinealidad. Los factores de inflación de la varianza (VIF) indican la posible existencia de colinealidad, o relación lineal entre las variables explicativas, cuyo efecto es añadir variabilidad a los coeficientes estimados, que se convierten en inestables y dejan de ser fiables. Un factor de inflación mayor que 30 indica un grave problema de multicolinealidad, lo que no ocurre en este caso. Un valor mayor que 15 muestra un cierto grado de relación lineal (multicolinealidad moderada). En este ejemplo se observa que no existe colinealidad entre las variables independientes (todos los valores menores que 2,7). Si existiera colinealidad entre las variables, sería necesario suprimir alguna de ellas.

```
VIF<-vif(Modelo.final)
VarInflFact<-paste("Factores de Inflacción de la Varianza para la estima de la colinealidad")
VIF
```

```
> VIF
             Ekman Plataforma.continental          Temperatura          Turbulencia
          1.938330               1.550211             2.375827             2.612731
```

Paso 18. Por último, se guardan todos los resultados en el archivo **Salida Cuadro XI.2.txt**.

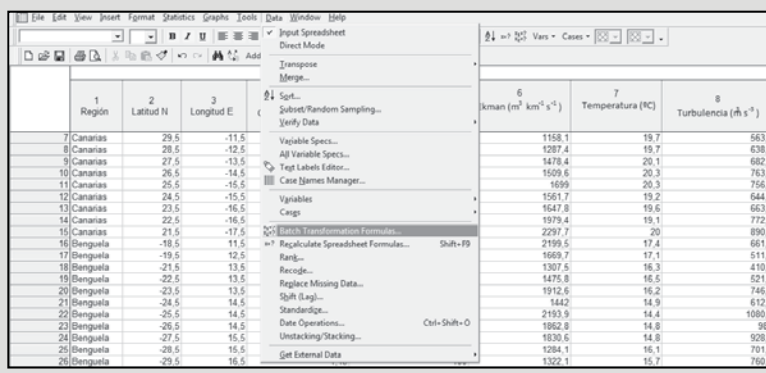
```
{
 sink("Salida Cuadro XI.2.txt")
 print(summary(Modelo.final))
 print(Shapiro)
 print(Lillie)
 print(BreuschPagan)
 print(DW)
 print(VarInflFact)
 print(VIF)
 sink()
}
```

CUADRO XI.2. *(Continuación)*[STATISTICA]

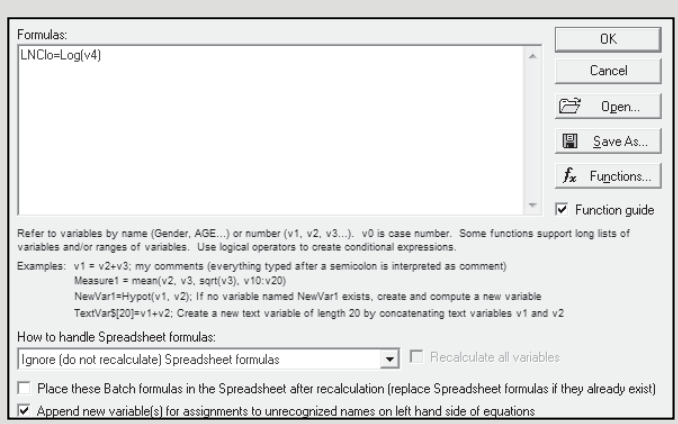
Aplicación con STATISTICA.

EJEMPLO. Se usa el mismo ejemplo que en la aplicación con R. En el archivo **Cuadro XI.2.sta** hay datos promedios a lo largo de varios años, para diferentes latitudes y longitudes en zonas de afloramiento, de clorofila *a* (mgm^{-3}), porcentaje que ocupa la plataforma continental, temperatura (ºC), transporte de Ekman ($m^3km^{-1}s^{-1}$, indicador de afloramiento) y turbulencia (m^3s^{-3}). El objetivo es determinar si es posible explicar la variación observada en la clorofila *a* en función de las otras variables, pero considerando todas las regiones.

Paso 1. Debido a que la relación entre la clorofila y alguna de las otras variables puede ser no lineal, es necesario hacer diferentes regresiones utilizando también las variables transformadas logarítmicamente. Para ello seleccionamos «Data» ⇨ «Batch Transformation Formulas…».

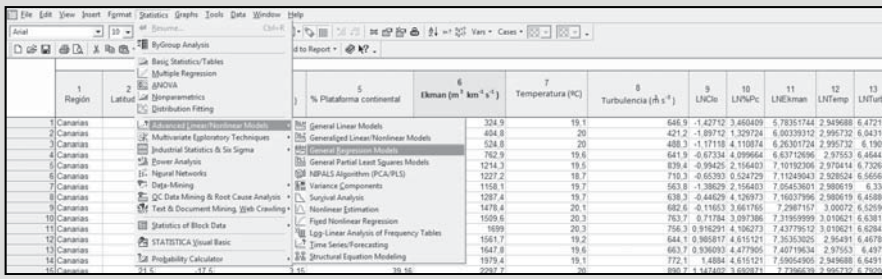

File Edit View Insert Format Statistics Graphs Tools Data Window Help
Input Spreadsheet
Direct Mode
Transpose
Merge...
Sort...
Subset/Random Sampling...
Verify Data
Variable Specs...
All Variable Specs...
Text Labels Editor...
Case Names Manager...
Variables
Cases
Batch Transformation Formulas...
Recalculate Spreadsheet Formulas... Shift+F9
Rank...
Recode...
Replace Missing Data...
Shift (Lag)...
Standardize...
Date Operations... Ctrl+Shift+O
Unstacking/Stacking...
Get External Data
1 Región
2 Latitud N
3 Longitud E
6 Temperatura (ºC)
8 Turbulencia
7 Canarias 29,5 -11,5 1158,1 19,7 563,8
8 Canarias 28,5 -12,5 1287,4 19,7 638,3
9 Canarias 27,5 -13,5 1478,4 20,1 682,6
10 Canarias 26,5 -14,5 1509,6 20,3 763,7
11 Canarias 25,5 -15,5 1699 20,3 756,3
12 Canarias 24,5 -15,5 1561,7 19,2 644,1
13 Canarias 23,5 -16,5 1647,8 19,6 663,7
14 Canarias 22,5 -16,5 1979,4 19,1 772,1
15 Canarias 21,5 -17,5 2297,7 20 890,7
16 Benguela -18,5 11,5 2199,5 17,4 661,7
17 Benguela -19,5 12,5 1669,7 17,1 511,7
18 Benguela -21,5 13,5 1307,5 16,3 410,9
19 Benguela -22,5 13,5 1475,8 16,5 521,5
20 Benguela -23,5 13,5 1912,6 16,2 746,5
21 Benguela -24,5 14,5 1442 14,9 612,4
22 Benguela -25,5 14,5 2193,9 14,4 1080,7
23 Benguela -26,5 14,5 1862,8 14,8 986
24 Benguela -27,5 15,5 1830,6 14,8 928,9
25 Benguela -28,5 15,5 1284,1 16,1 701,4
26 Benguela -29,5 16,5 1322,1 15,7 760,4

Paso 2. Se introduce la siguiente expresión «LNClo=Log(v4)», con la cual creamos una nueva variable que es el logaritmo de la clorofila. Se repite el proceso con las otras variables (porcentaje de la plataforma continental, temperatura, transporte de Ekman y turbulencia). En el caso de la plataforma continental hay que realizar la transformación sumando 1, ya que hay valores 0, «LN%Pc=Log(v5+1)»,

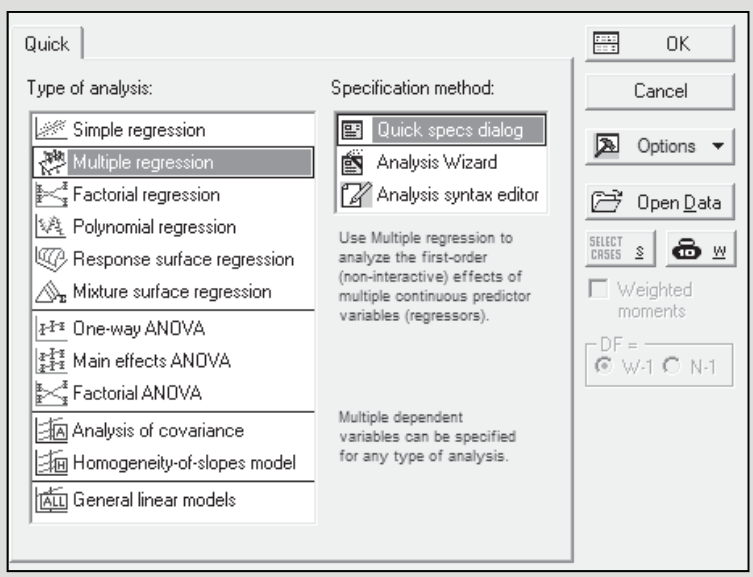

Formulas:
LNClo=Log(v4)
OK
Cancel
Open...
Save As...
Functions...
Function guide
Refer to variables by name (Gender, AGE...) or number (v1, v2, v3...). v0 is case number. Some functions support long lists of variables and/or ranges of variables. Use logical operators to create conditional expressions.
Examples: v1 = v2+v3; my comments (everything typed after a semicolon is interpreted as comment)
Measure1 = mean(v2, v3, sqrt(v3), v10:v20)
NewVar1=Hypot(v1, v2); If no variable named NewVar1 exists, create and compute a new variable
TextVar$[20]=v1+v2; Create a new text variable of length 20 by concatenating text variables v1 and v2
How to handle Spreadsheet formulas:
Ignore (do not recalculate) Spreadsheet formulas
Recalculate all variables
Place these Batch formulas in the Spreadsheet after recalculation (replace Spreadsheet formulas if they already exist)
Append new variable(s) for assignments to unrecognized names on left hand side of equations

CUADRO XI.2. *(Continuación)*[STATISTICA]

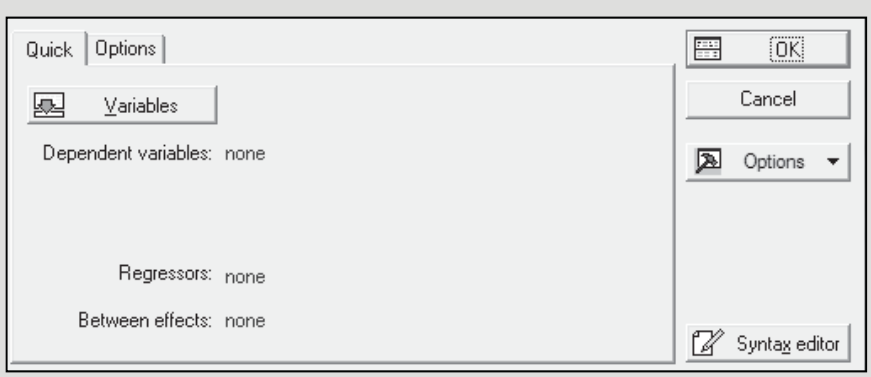
Paso 3. A continuación se selecciona «Statistics» ⇨ «Advanced Linear/Nonlinear Models» ⇨ «General Regression Models».



Paso 4. En la ventana que aparece se selecciona «Multiple regression» y «Quick specs dialog».

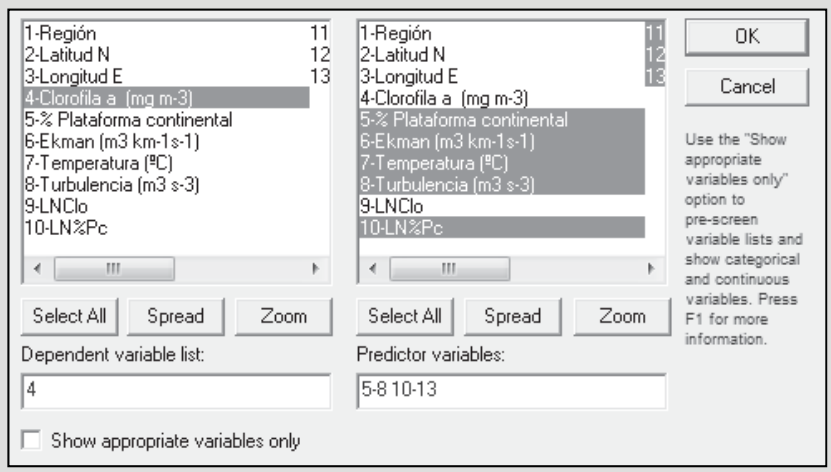


Paso 5. Se pulsa en «Variables».

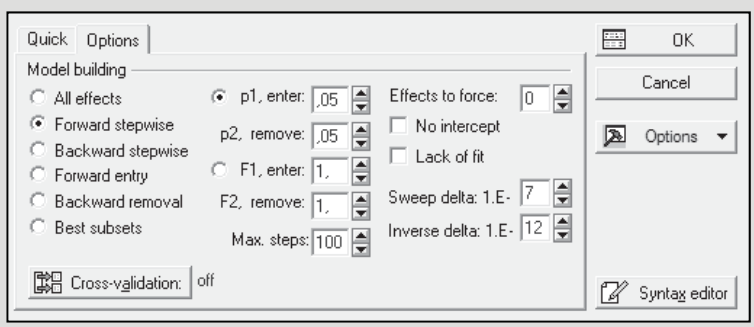


CUADRO XI.2. *(Continuación)*[STATISTICA]

Paso 6. Se selecciona la variable dependiente, en este caso la clorofila, y las variables independientes, que serían el resto de variables. En primer lugar realizaremos el análisis con la variable dependiente no transformada logarítmicamente y como independientes, se incluyen todas las variables, tanto transformadas como no transformadas, lógicamente a excepción de la transformación logarítmica de la clorofila.



Paso 7. En la ventana del paso 5, en la pestaña «Options» se observa que hay varias formas de construir el modelo de regresión



- *All effects*. Nos muestra la ecuación con todas las variables independientes. Es decir, introduce todas las variables en la ecuación aunque no sean significativas.

- *Forward stepwise.* En cada paso se introduce la variable independiente que no se encuentre ya en la ecuación y que tenga el mayor valor de *F*, siempre que sea suficientemente grande. Las variables ya introducidas en la ecuación de regresión se eliminan de ella si su valor *F* llega a ser pequeño. El método termina cuando ya no hay más variables para incluir o excluir. Este método es, junto con el anterior, el más usado

CUADRO XI.2. *(Continuación)*[STATISTICA]

- *Backward stepwise*. Similar al anterior, aunque ahora se parte de todas las variables en la ecuación, y se van eliminando aquellas cuyo valor F es más pequeño (siempre que no sea suficientemente grande). Eventualmente puede volver a introducirse una variable ya eliminada si su valor F se hace grande.

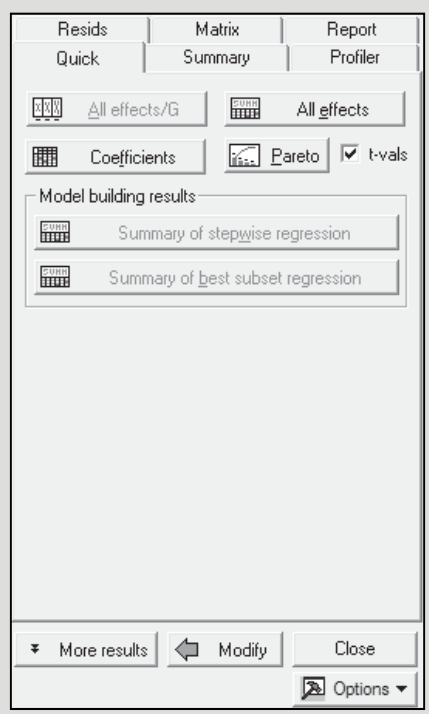
- *Forward entry*. Similar a *Forward stepwise,* aunque no se eliminan variables ya introducidas.

- *Backward removal*. Similar a *Backward stepwise,* aunque no se vuelve a introducir una variable que ya ha sido eliminada.

- *Best subsets*. La introducción o eliminación de variables paso a paso no garantiza la obtención del mejor subconjunto posible de variables. Los métodos "best subsets" exploran todos los subconjuntos posibles con un número reducido de variables, el cual debe ser definido previamente, y tratan de encontrar el mejor subconjunto mediante criterios estadísticos.

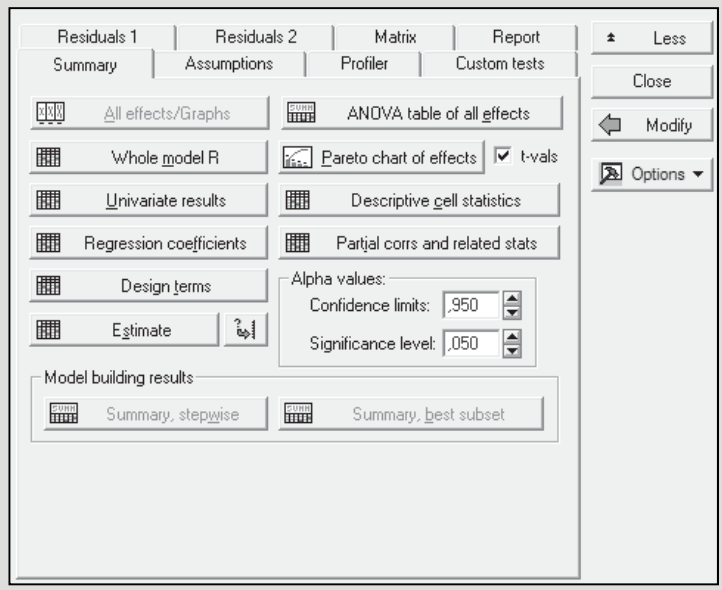
Vamos a trabajar con la opción por pasos hacia adelante, es decir, «Forward stepwise». Sale un mensaje de aviso en el que pulsamos «Aceptar».

Paso 8. En la nueva ventana que aparece se pulsa en «More results».



CUADRO XI.2. *(Continuación)* STATISTICA

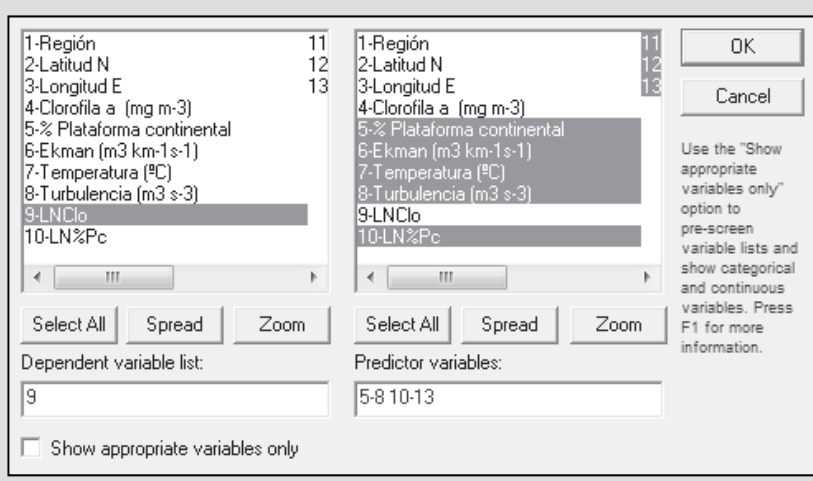
Paso 9. En la pestaña «Summary» se pulsa en «Whole model R».



Paso 10. Se observa que el modelo es significativo ($F_{4,\ 56}$ = 33,01, p < 0,001, r^2 = 0,7).

	Test of SS Whole Model vs. SS Residual (Cuadro XI.2.sta)										
Dependnt Variable	Multiple R	Multiple R^2	Adjusted R^2	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Clorofila a (mg m^{-3})	0,837990	0,702228	0,680958	69,50203	4	17,37551	29,47160	56	0,526279	33,01580	0,000000

Paso 11. Se repite el proceso pero esta vez con la variable dependiente transformada logarítmicamente y, de nuevo, como independientes se incluyen todas las variables, tanto transformadas como no transformadas. De nuevo se usa la opción «Forward stepwise» y en el mensaje de aviso pulsamos «Aceptar».



CUADRO XI.2. *(Continuación)*[STATISTICA]

Paso 12. El nuevo modelo es mejor que el anterior y, por tanto, nos quedamos con este último ($F_{4, 56}$ = 44,002, p < 0,001, r^2 = 0,76).

	Test of SS Whole Model vs. SS Residual (Cuadro XI.2.sta)										
Dependnt Variable	Multiple R	Multiple R²	Adjusted R²	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
LNClo	0,870995	0,758632	0,741391	30,83745	4	7,709363	9,811333	56	0,175202	44,00262	0,000000

A veces la opción por pasos no funciona bien y no escoge las variables que dan lugar al mejor modelo. Por ello, es conveniente realizar pruebas con diferentes combinaciones de las variables transformadas y no transformadas para obtener el mejor modelo usando la opción «All effects» en el paso 7, y así incluir todas las variables independientes que se indiquen. En este caso, se hicieron todas las pruebas posibles y no se obtuvo un modelo mejor que el anterior.

Paso 13. En la ventana del paso 9 se pulsa en «Univariate results» y se obtienen las variables que se incluyen en el modelo, que al ser la opción por pasos, son todas significativas (p < 0,05).

	Univariate Results for Each DV (Cuadro XI.2.sta) Forward stepwise solution Effective hypothesis decomposition				
Effect	Degr. of Freedom	LNClo SS	LNClo MS	LNClo F	LNClo p
Intercept	1	2,57570	2,57570	14,70128	0,000322
% Plataforma continental	1	6,54051	6,54051	37,33119	0,000000
Ekman ($m^3\,km^{-1}s^{-1}$)	0				
Temperatura (ºC)	1	4,59541	4,59541	26,22915	0,000004
Turbulencia ($m^3\,s^{-3}$)	1	2,22190	2,22190	12,68190	0,000762
LN%Pc	0				
LNEkman	1	12,94703	12,94703	73,89759	0,000000
LNTemp	0				
LNTurb	0				
Error	56	9,81133	0,17520		
Total	60	40,64878			

Paso 14. En la ventana del paso 9 se pulsa en «Regression coefficients» y se obtienen los coeficientes de la regresión.

	Parameter Estimates (Cuadro XI.2.sta) Sigma-restricted parameterization										
Effect	Comment (B/Z/P)	LNClo Param.	LNClo Std.Err	LNClo t	LNClo p	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt	LNClo Beta (ß)	LNClo St.Err.ß	-95,00% Cnf.Lmt	+95,00% Cnf.Lmt
Intercept		-3,77575	0,984751	-3,83422	0,000322	-5,74845	-1,80306				
% Plataforma continental		0,01043	0,001707	6,10993	0,000000	0,00701	0,01385	0,424964	0,069553	0,285632	0,564295
Ekman ($m^3\,km^{-1}s^{-1}$)	Pooled										
Temperatura (ºC)		-0,16170	0,031573	-5,12144	0,000004	-0,22495	-0,09845	-0,447635	0,087404	-0,622727	-0,272543
Turbulencia ($m^3\,s^{-3}$)		-0,00086	0,000241	-3,56117	0,000762	-0,00134	-0,00038	-0,314469	0,088305	-0,491365	-0,137573
LN%Pc	Pooled										
LNEkman		0,99363	0,115587	8,59637	0,000000	0,76208	1,22518	0,600736	0,069883	0,460745	0,740728
LNTemp	Pooled										
LNTurb	Pooled										

CUADRO XI.2. *(Continuación)*[STATISTICA]

Por lo tanto, las variables porcentaje que ocupa la plataforma continental (*PT*, %), temperatura (*T*, ºC), la transformación logarítmica del transporte de Ekman (*Ek*, $m^3\ km^{-1}s^{-1}$) y la turbulencia (*Tb*, $m^3\ s^{-3}$) explican el 76% de la variación observada en la clorofila a (*Chl*, mg m^{-3}) transformada logarítmicamente, por medio de la siguiente regresión.

$$\ln(Chl) = -3{,}775 + 0{,}0104*PT - 0{,}1617*T - 0{,}00086*Tb + 0{,}9936*\ln(Ek)$$

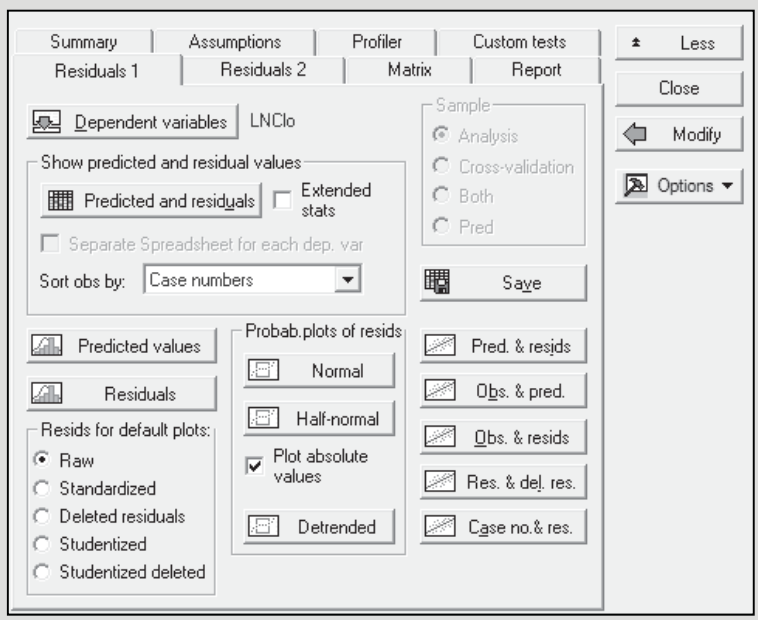
Paso 15. En la ventana del paso 9 al pulsar «Summary stepwise» aparece la siguiente tabla donde se muestra como fue cada uno de los pasos de la regresión y, por tanto, se confirma que la variable más importante es el transporte de Ekman, después la plataforma continental, temperatura y, por último, la turbulencia.

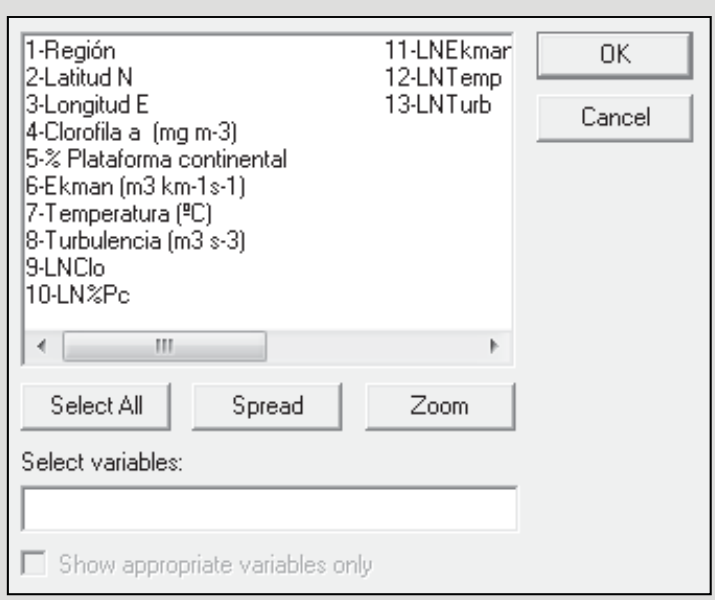
	Summary of stepwise regression; variable: LNClo (Cuadro XI.2.sta) Forward stepwise P to enter: ,05, P to remove: ,05						
Effect	Steps	Degr. of Freedom	F to remove	P to remove	F to enter	P to enter	Effect status
% Plataforma continental	Step Number 1	1			33,19324	0,000000	Out
Ekman ($m^3\ km^{-1}s^{-1}$)		1			41,12176	0,000000	Out
Temperatura (ºC)		1			4,38503	0,040561	Out
Turbulencia ($m^3\ s^{-3}$)		1			0,66121	0,419403	Out
LN%Pc		1			26,55048	0,000003	Out
LNEkman		1			55,89552	0,000000	Entered
LNTemp		1			3,58531	0,063199	Out
LNTurb		1			1,08483	0,301869	Out
LNEkman	Step Number 2	1	55,89552	0,000000			In
Ekman ($m^3\ km^{-1}s^{-1}$)		1			0,57222	0,452439	Out
Temperatura (ºC)		1			9,04960	0,003882	Out
Turbulencia ($m^3\ s^{-3}$)		1			0,04510	0,832573	Out
LN%Pc		1			20,90651	0,000026	Out
% Plataforma continental		1			25,97189	0,000004	Entered
LNTemp		1			7,23745	0,009310	Out
LNTurb		1			0,05600	0,813772	Out
LNEkman	Step Number 3	1	46,64969	0,000000			In
% Plataforma continental		1	25,97189	0,000004			In
Temperatura (ºC)		1			11,29409	0,001393	Entered
Turbulencia ($m^3\ s^{-3}$)		1			0,04264	0,837139	Out
LN%Pc		1			0,24614	0,621715	Out
Ekman ($m^3\ km^{-1}s^{-1}$)		1			0,41191	0,523573	Out
LNTemp		1			9,60737	0,003009	Out
LNTurb		1			0,53994	0,465470	Out
LNEkman	Step Number 4	1	55,78672	0,000000			In
% Plataforma continental		1	28,53046	0,000002			In
Temperatura (ºC)		1	11,29409	0,001393			In
Turbulencia ($m^3\ s^{-3}$)		1			12,68190	0,000762	Entered
LN%Pc		1			1,31836	0,255769	Out
Ekman ($m^3\ km^{-1}s^{-1}$)		1			0,16131	0,689478	Out
LNTemp		1			9,18642	0,003687	Out
LNTurb		1			12,63998	0,000776	Out
LNEkman	Step Number 5	1	73,89759	0,000000			In
% Plataforma continental		1	37,33119	0,000000			In
Temperatura (ºC)		1	26,22915	0,000004			In
Turbulencia ($m^3\ s^{-3}$)		1	12,68190	0,000762			In
LN%Pc		1			2,32257	0,133239	Out
Ekman ($m^3\ km^{-1}s^{-1}$)		1			0,23605	0,629002	Out
LNTemp		1			2,70246	0,105899	Out
LNTurb		1			0,42368	0,517817	Out

CUADRO XI.2. *(Continuación)*[STATISTICA]

A continuación veremos si se cumplen los requisitos de Normalidad y homocedasticidad de los residuos, y si existe colinealidad entre las variables independientes. No existe la opción de testar la autocorrelación con el test de Durbin-Watson.

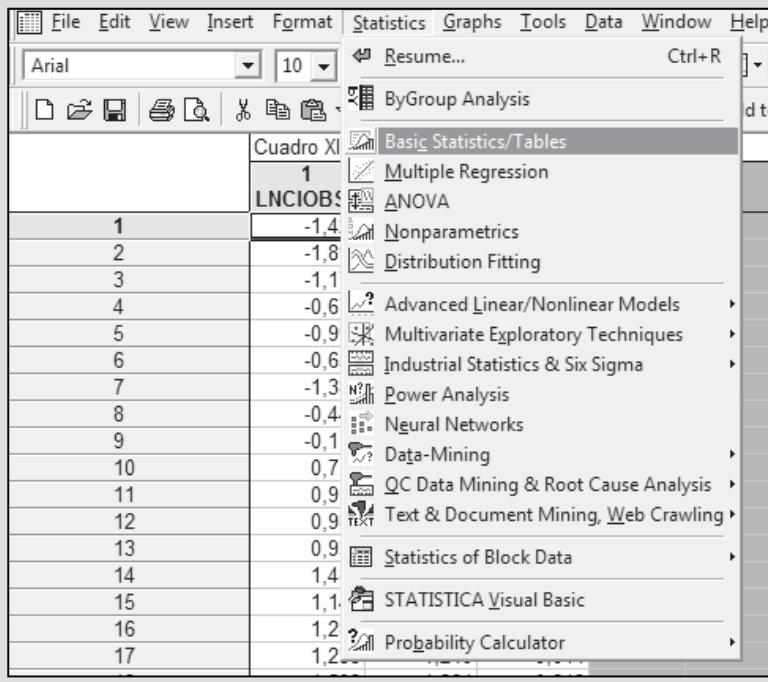
Paso 16. Distribución normal de los residuos. En la pestaña «Residuals 1» se pulsa en «Save», y en la ventana que aparece se pulsa «OK» sin seleccionar nada.



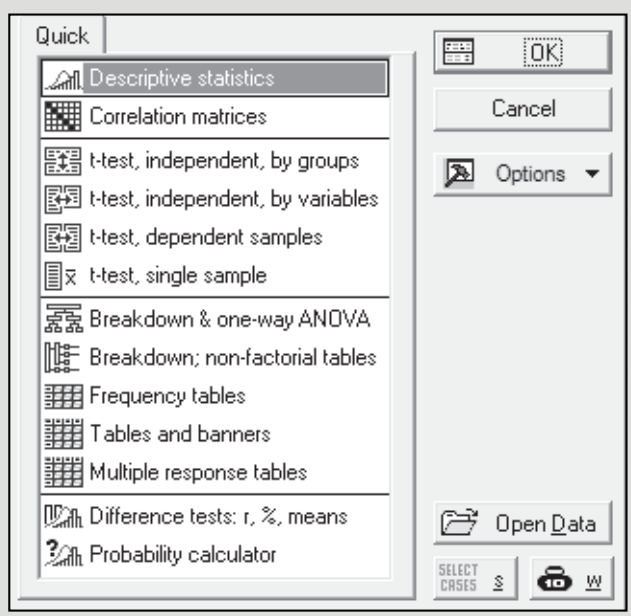


CUADRO XI.2. *(Continuación)*[STATISTICA]

Aparece una tabla donde la tercera columna son los residuos. Para ver si los residuos tienen distribución Normal se selecciona «Statistics» ⇨ «Basic Statistics/Tables».

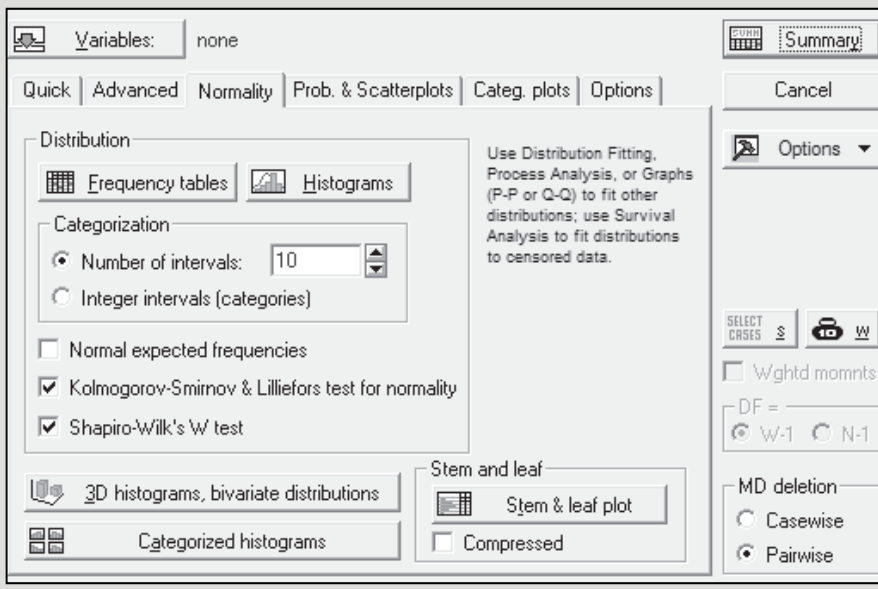


En la siguiente ventana seleccionamos «Descriptive statistics».

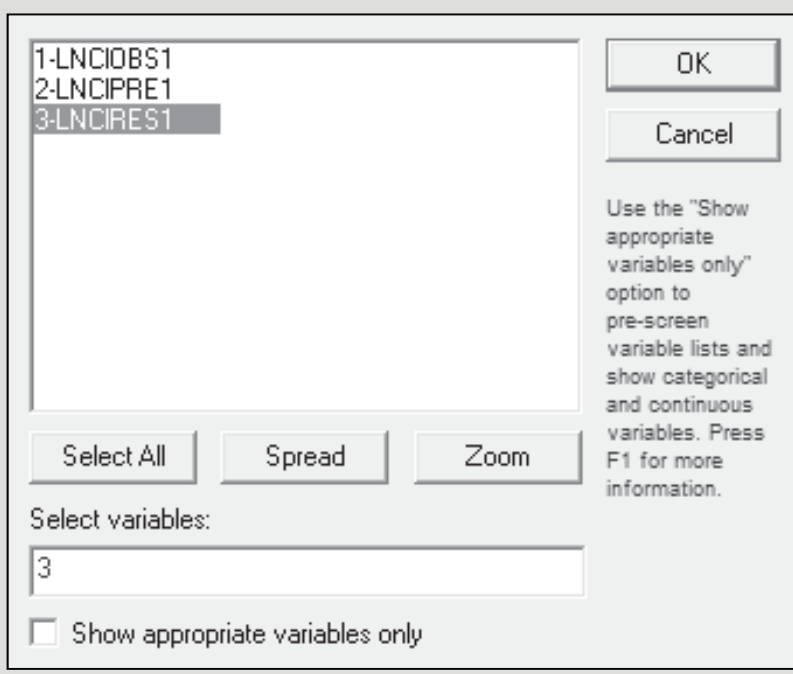


CUADRO XI.2. *(Continuación)* STATISTICA

En la siguiente ventana, en la pestaña «Normality», se seleccionan los tests «Kolmogorov-Smirnov & Lilliefors test for normality» y «Shapiro-Wilk´s W test», y pulsamos en el botón «Variables»

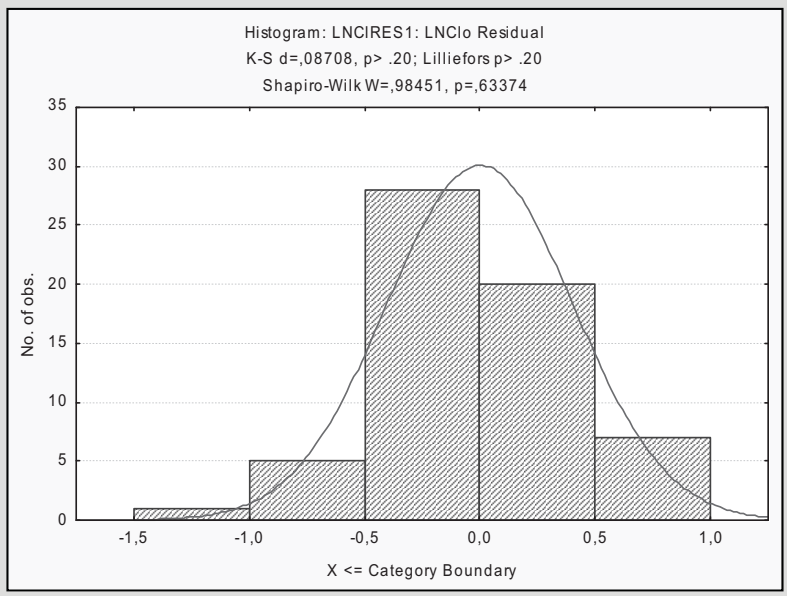


A continuación se seleccionan los residuos «LNCIRES1».

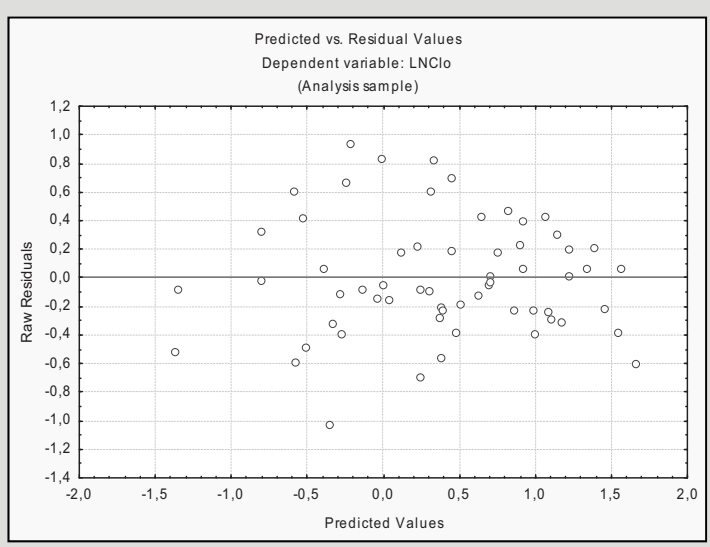


CUADRO XI.2. *(Continuación)*[STATISTICA]

Por último, pulsamos el botón «Histograms» y aparece una gráfica donde se observa que los residuos tienen distribución Normal, ya que tanto en el test K-S con la corrección de Lilliefors (p > 0,2), como en el test Shapiro-Wilk (p = 0,663), la probabilidad es mayor que 0,05.

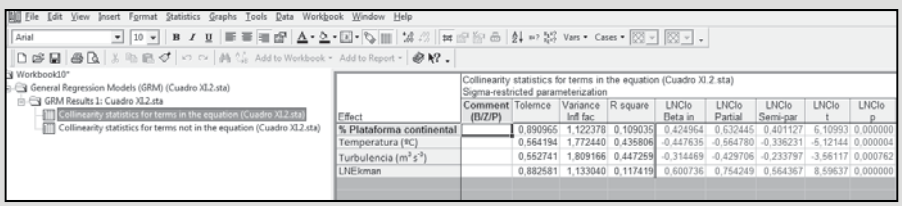


Paso 17. Homocedasticidad de los residuos. En la pestaña «Residuals 1» del paso 16, se pulsa en «Pred. & resid». La gráfica muestra que los residuos son pequeños y la variabilidad se mantiene más o menos constante a lo largo de todo el rango de los valores pronosticados tipificados. Por tanto, aceptamos que existe homocedasticidad en los residuos.



CUADRO XI.2. *(Continuación)*[STATISTICA]

Paso 18. Colinealidad. En la ventana del paso 9 al pulsar «Partial corrs and related stats» aparecen los indicadores de colinealidad. En el lado izquierdo hay que seleccionar «Collinearity statistics for terms in the equation», para que nos salgan los índices de colinealidad de los coeficientes que se incluyen en el modelo de regresión, es decir, seleccionando la línea que está resaltada en el lado izquierdo de la pantalla.


File Edit View Insert Format Statistics Graphs Tools Data Workbook Window Help
Workbook10*
General Regression Models (GRM) (Cuadro XI.2.sta)
GRM Results 1: Cuadro XI.2.sta
Collinearity statistics for terms in the equation (Cuadro XI.2.sta)
Collinearity statistics for terms not in the equation (Cuadro XI.2.sta)
Collinearity statistics for terms in the equation (Cuadro XI.2.sta)
Sigma-restricted parameterization
Effect | Comment (B/Z/P) | Tolernce | Variance Infl fac | R square | LNClo Beta in | LNClo Partial | LNClo Semi-par | LNClo t | LNClo p
% Plataforma continental | | 0,890965 | 1,122378 | 0,109035 | 0,424964 | 0,632445 | 0,401127 | 6,10993 | 0,000000
Temperatura (ºC) | | 0,564194 | 1,772440 | 0,435806 | -0,447635 | -0,564780 | -0,336231 | -5,12144 | 0,000004
Turbulencia (m³s⁻³) | | 0,552741 | 1,809166 | 0,447259 | -0,314469 | -0,429706 | -0,233797 | -3,56117 | 0,000762
LNEkman | | 0,882581 | 1,133040 | 0,117419 | 0,600736 | 0,754249 | 0,564367 | 8,59637 | 0,000000

Los índices que se utilizan para estudiar la colinealidad son la «Tolerancia» y «FIV» (Factor de Inflación de la Varianza, el inverso de la tolerancia). La tolerancia es la proporción de la varianza de cada variable independiente que no es explicada por las restantes variables independientes, y se obtiene restando de la unidad el coeficiente de determinación r^2 de una regresión múltiple realizada con cada una de las variables explicativas como dependiente de las restantes. Aunque no existen reglas fijas, una tolerancia muy pequeña, o un FIV alto (por ejemplo FIV > 15), muestra un posible problema de multicolinealidad.

La tolerancia para la variable plataforma continental es 0,89, para la temperatura 0,56, para la turbulencia 0,5 y para el transporte de Ekman 0,88. Como todos los valores de tolerancia son superiores al 50% (> 0,5), no existe colinealidad. En el caso de que existiese colinealidad entre variables, sería necesario eliminar alguna variable del modelo.

La tabla anterior también muestra las correlaciones parciales, donde de nuevo podemos ver que la variable más importante es el transporte de Ekman (0,75), después el tamaño de la plataforma continental (0,63), la temperatura (-0,56) y, por último, la turbulencia (-0,43).

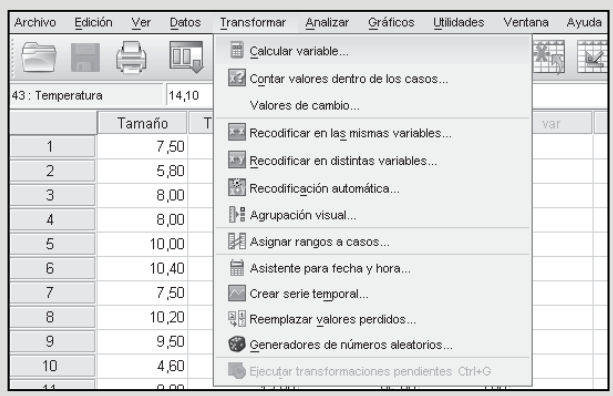
En resumen, gran parte de la variación observada en la concentración de clorofila entre las diferentes regiones (76%) se puede explicar por un modelo en el cual el afloramiento de las masas de agua (transporte de Ekman) y el tamaño de la plataforma continental tienen una relación positiva con la clorofila, mientras que la turbulencia y la temperatura tienen una relación negativa.

CUADRO XI.2. *(Continuación)*[SPSS]

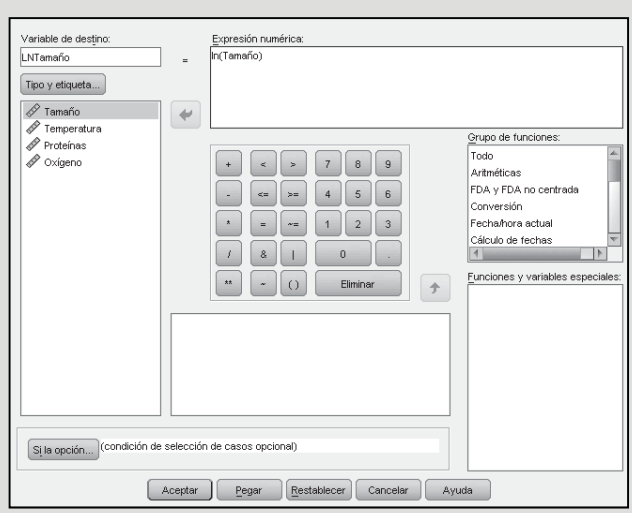
Aplicación con SPSS.

EJEMPLO. En el archivo **Cuadro XI.2.sav** se muestran datos sobre el tamaño (en mm) de la larva y el contenido en proteínas de la larva (en µg por larva) de una especie de pez, recolectados en distintas estaciones de muestreo donde se midió la temperatura (ºC) y la concentración de oxigeno del agua (mg ml^{-1}). Se quiere determinar si el contenido en proteínas de la larva depende del tamaño de la larva, concentración de oxígeno y/o de la temperatura, y encontrar la función que los relaciona.

Paso 1. Debido a que la relación entre el contenido en proteínas y las variables independientes puede que no sea lineal, es necesario hacer diferentes regresiones utilizando también las variables transformadas logarítmicamente. Para transformar las variables seleccionamos «Transformar» ⇨ «Calcular variable…».

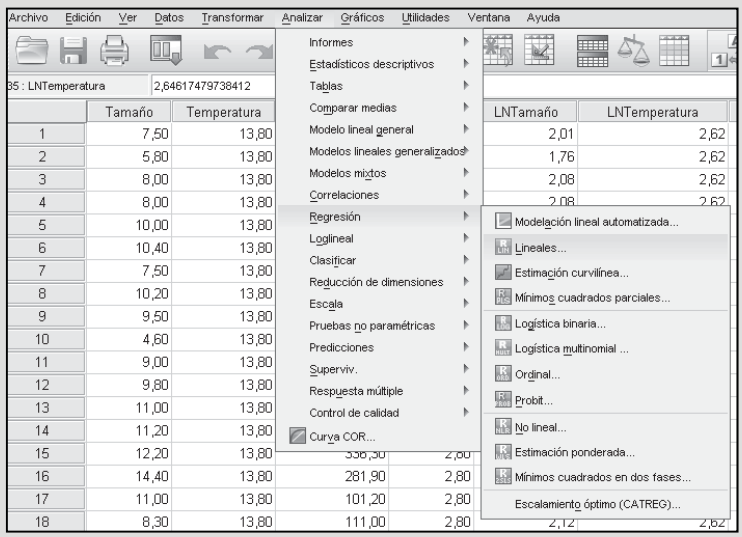


Paso 2. En «Variable de destino:» creamos una nueva variable «LNTamaño» y en «Expresión numérica:» ponemos la transformación logarítmica «ln(Tamaño)». Este proceso hay que hacerlo con las otras variables: temperatura, proteínas y oxígeno.

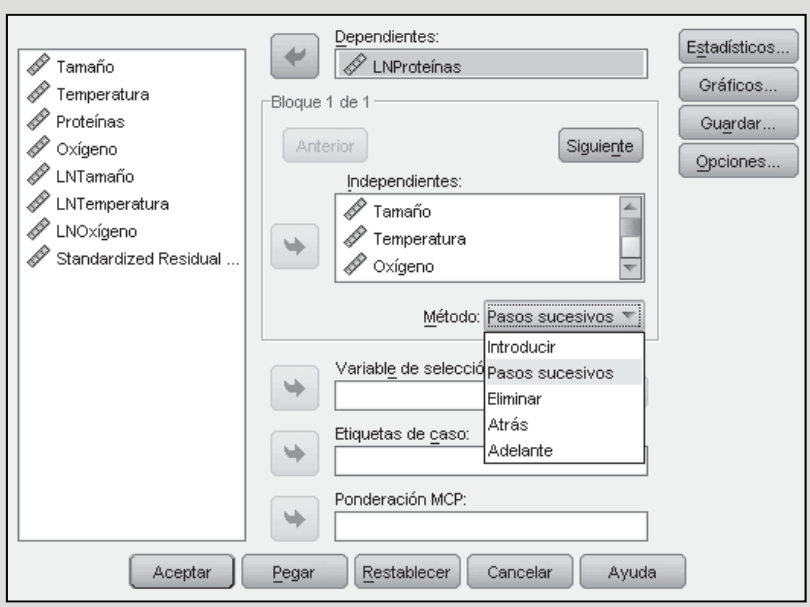


CUADRO XI.2. *(Continuación)*[SPSS]

Paso 3. Posteriormente entramos en «Analizar» ⇨ «Regresión» ⇨ «Lineales... ».



Paso 4. En primer lugar realizamos el análisis con la variable dependiente transformada «LNProteínas» y como variables independientes se incluye el resto de variables, tanto transformadas como no transformadas.



En «Método:» se puede elegir entre distintas formas de introducir las variables en la ecuación:

CUADRO XI.2. *(Continuación)*[SPSS]

- *Introducir*. Nos muestra la ecuación con todas las variables independientes. Es decir, introduce todas las variables en la ecuación aunque no sean significativas.

- *Pasos sucesivos.* En cada paso se introduce la variable independiente que no se encuentre ya en la ecuación y que tenga una probabilidad para *F* suficientemente pequeña. Las variables ya introducidas en la ecuación de regresión se eliminan de ella si su probabilidad para *F* llega a ser grande. Es decir, sólo introduce en la ecuación las variables que son significativas. Es junto con el anterior el más usado.

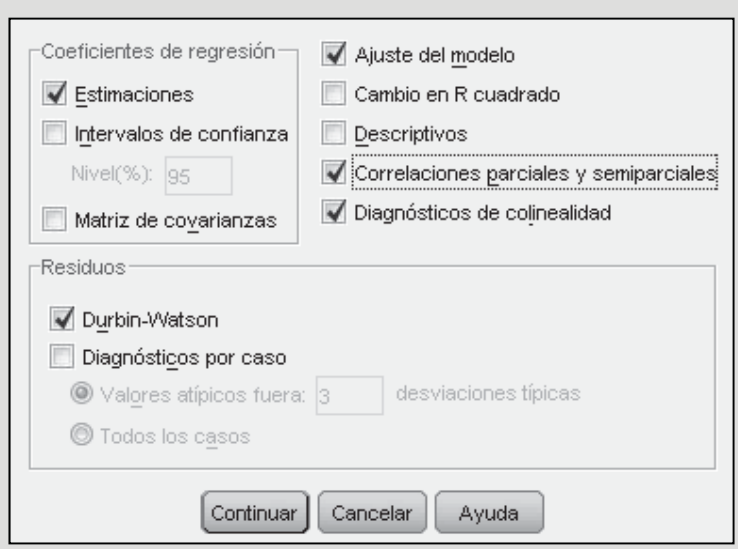
- *Eliminar.* Procedimiento para la selección de variables en el que las variables de un bloque se eliminan en un solo paso.

- *Atrás*. Procedimiento de selección de variables en el que se introducen todas las variables en la ecuación y después se van excluyendo una tras otra. Aquella variable que tenga la menor correlación parcial con la variable dependiente será la primera en ser considerada para su exclusión. El procedimiento termina cuando ya no quedan en la ecuación variables que satisfagan el criterio de exclusión.

- *Adelante*. Es el proceso contrario al anterior. Va introduciendo primero las variables con mayor significación, hasta que no queda ninguna variable significativa por introducir.

Vamos a trabajar con la opción «Pasos sucesivos», para que se incluyan en el modelo solo las variables que son significativas.

Paso 5. Se siguen los pasos 12 a 15 del Cuadro XI.1 con la excepción de que en «Estadísticos» se marca a mayores «Diagnósticos de colinealidad» y «Correlaciones parciales y semiparciales».



CUADRO XI.2. *(Continuación)*[SPSS]

Paso 6. Los resultados del análisis son los siguientes, de los cuales la mayoría de los parámetros ya se explicaron en el Cuadro XI.1.

En la tabla «Resumen del modelo» el r^2 mayor es el del último modelo (0,825) y, por tanto, el ajuste es mejor (siempre será así cuando se añaden variables a un modelo).

Resumen del modelo[d]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,858[a]	,737	,734	,62496	
2	,901[b]	,811	,807	,53244	
3	,909[c]	,825	,819	,51452	1,947

a. Variables predictoras: (Constante), LNTamaño

b. Variables predictoras: (Constante), LNTamaño, LNTemperatura

c. Variables predictoras: (Constante), LNTamaño, LNTemperatura, Temperatura

d. Variable dependiente: LNProteínas

La tabla ANOVA muestra que todos los modelos son significativos, con $p < 0{,}001$.

ANOVA[d]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	96,085	1	96,085	246,011	,000[a]
	Residual	34,370	88	,391		
	Total	130,455	89			
2	Regresión	105,791	2	52,895	186,583	,000[b]
	Residual	24,664	87	,283		
	Total	130,455	89			
3	Regresión	107,688	3	35,896	135,595	,000[c]
	Residual	22,767	86	,265		
	Total	130,455	89			

a. Variables predictoras: (Constante), LNTamaño

b. Variables predictoras: (Constante), LNTamaño, LNTemperatura

c. Variables predictoras: (Constante), LNTamaño, LNTemperatura, Temperatura

d. Variable dependiente: LNProteínas

En la tabla «Coeficientes» vemos cuales son las variables independientes que se incluyen en cada uno de los modelos. El tercer modelo incluye como variables independientes a «Temperatura» y «LNTemperatura», con lo cual es seguro que existirán problemas de colinealidad. Lógicamente, en todos los modelos las variables independientes son significativas, con $p < 0{,}05$.

CUADRO XI.2. *(Continuación)*[SPSS]

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados			Correlaciones			Estadísticos de colinealidad	
		B	Error típ.	Beta	t	Sig.	Orden cero	Parcial	Semiparcial	Tolerancia	FIV
1	(Constante)	-1,567	,354		-4,434	,000					
	LNTamaño	2,481	,158	,858	15,685	,000	,858	,858	,858	1,000	1,000
2	(Constante)	11,876	2,317		5,125	,000					
	LNTamaño	2,649	,138	,916	19,226	,000	,858	,900	,896	,956	1,046
	LNTemperatura	-5,127	,876	-,279	-5,851	,000	-,088	-,531	-,273	,956	1,046
3	(Constante)	217,83	76,967		2,830	,006					
	LNTamaño	2,507	,143	,867	17,485	,000	,858	,883	,788	,825	1,212
	LNTemperatura	-125,8	45,084	-6,844	-2,790	,006	-,088	-,288	-,126	,000	2964,1
	Temperatura	8,058	3,010	6,576	2,677	,009	-,079	,277	,121	,000	2973,5

a. Variable dependiente: LNProteínas

Paso 7. Colinealidad. En la tabla «Coeficientes» mostrada en el paso anterior se muestran dos estadísticos denominados «Tolerancia» y «FIV» (Factor de Inflación de la Varianza, el inverso de la tolerancia) que se utilizan para estudiar la colinealidad. La tolerancia es la proporción de la varianza de cada variable independiente que no es explicada por las restantes variables independientes, y se obtiene restando de la unidad el coeficiente de determinación r^2 de una regresión múltiple realizada con cada una de las variables explicativas como dependiente de las restantes. Aunque no existen reglas fijas, una tolerancia muy pequeña o un FIV alto (por ejemplo FIV > 15), muestra una variable que casi es combinación lineal de las restantes, lo que indicaría un posible problema de colinealidad. En el tercer modelo la tolerancia es muy baja (vale 0) y el FIV muy alto para «Temperatura» (2973,5) y «LNTemperatura» (2964,1). Por tanto, es necesario eliminar una de las dos variables. Esto quiere decir que el modelo correcto es el dos, el cual solo tiene a «LNTemperatura».

En la tabla «Diagnósticos de colinealidad» se muestran otros resultados para estudiar este problema. En ellos se muestran los «Autovalores», que indican las dimensiones subyacentes: si hay varios autovalores muy próximos a cero, ello indicaría la presencia de colinealidad. Para su mejor valoración se calcula el «Indice de condición», o raíz cuadrada del cociente entre el mayor autovalor y los restantes. Un índice de condición mayor que 15 indica la posible presencia de colinealidad, y si es mayor que 30 indica un serio problema. Sin embargo, dado que la constante del modelo añade una dimensión más, debe matizarse con las «Proporciones de la varianza» de cada coeficiente de regresión que explica cada dimensión (en general cada dimensión explica un porcentaje alto de solamente una variable explicativa). Si una dimensión con índice de condición elevado explica un porcentaje alto de dos o más variables explicativas (la constante no cuenta), existe un serio problema de colinealidad. En nuestro ejemplo se confirma que hay colinealidad en el tercer modelo, ya que hay dos variables, «Temperatura» y «LNTemperatura», que ambas se explican con un porcentaje alto (1). Por lo tanto, como se dijo anteriormente hay que eliminar «Temperatura».

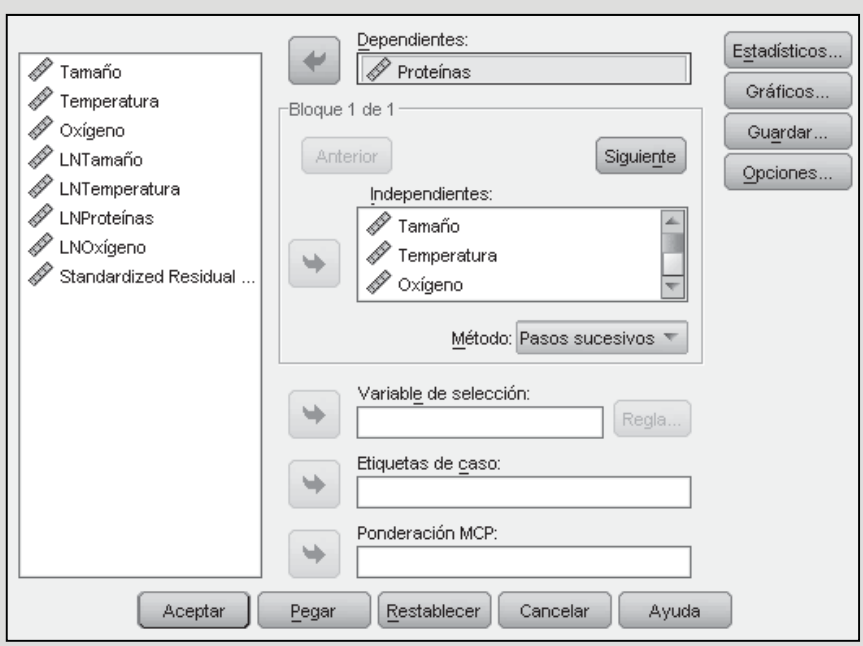
CUADRO XI.2. *(Continuación)*[SPSS]

Diagnósticos de colinealidad[a]

Modelo	Dimensión		dimension2					
					Proporciones de la varianza			
			Autovalores	Índice de condición	(Constante)	LNTamaño	LNTemperatura	Temperatura
dimension0	1	dimension1 1	1,982	1,000	,01	,01		
		2	,018	10,639	,99	,99		
	2	dimension1 1	2,977	1,000	,00	,00	,00	
		2	,023	11,426	,00	,97	,00	
		3	,000	101,498	1,00	,02	1,00	
	3	dimension1 1	3,973	1,000	,00	,00	,00	,00
		2	,025	12,583	,00	,86	,00	,00
		3	,002	42,186	,00	,00	,00	,00
		4	1,297E-7	5535,029	1,00	,14	1,00	1,00

a. Variable dependiente: LNProteínas

Paso 8. Se repite el análisis pero con la variable dependiente sin transformar «Proteínas», por si se obtiene un mejor modelo y, al igual que antes, como variables dependientes se incluye de nuevo el resto de variables, tanto transformadas como no transformadas.



Paso 9. En la tabla «Resumen del modelo» el r^2 mayor es 0,58, inferior al que se obtenía en el segundo modelo cuando la variable dependiente es «LNProteínas» (0,81, véase paso 6).

CUADRO XI.2. *(Continuación)*[SPSS]

Resumen del modelo[d]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,631[a]	,398	,391	64,48997	
2	,733[b]	,537	,526	56,89617	
3	,762[c]	,581	,566	54,45632	1,490

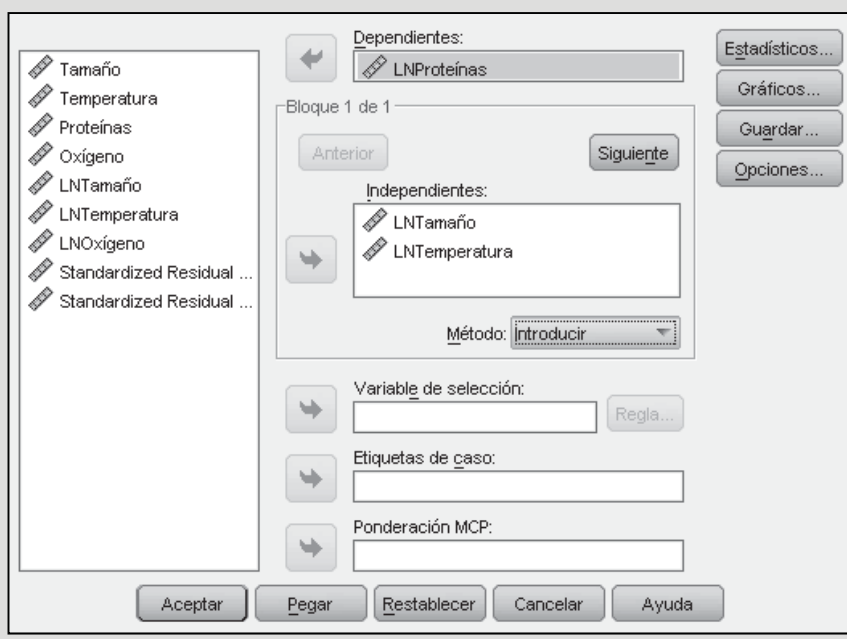
a. Variables predictoras: (Constante), Tamaño

b. Variables predictoras: (Constante), Tamaño, Temperatura

c. Variables predictoras: (Constante), Tamaño, Temperatura, LNTamaño

d. Variable dependiente: Proteínas

Paso 10. Se repite de nuevo el análisis, pero solo incluyendo las variables que dan el mejor modelo, es decir, como variable dependiente «LNProteínas» y como independientes «LNTamaño» y «LNTemperatura». En «Método» se escoge la opción «Introducir», ya que se quiere que se consideren todas las variables independientes.



Paso 11. Los resultados muestran que el modelo explica un 81% de la varianza observada en las proteínas de la larva y es significativo con $p < 0{,}001$. En la tabla «Coeficientes» en «Sig.» se observa que, tanto la constante como las dos variables independientes son significativas con una $p < 0{,}001$. Las correlaciones parciales muestran que el tamaño tiene una mayor relación (0,9) que la temperatura (-0,53) con la proteína de la larva.

CUADRO XI.2. *(Continuación)*[SPSS]

La forma de expresar nuestros resultados sería la siguiente: existe una relación significativa entre las proteínas de la larva con el tamaño de la larva y la temperatura (r^2 = 0,81, $F_{2,87}$ = 186,58, p < 0,001), que se describe por medio de la siguiente ecuación:

ln(*Proteínas*) = 11,876 + 2,649*ln(*Tamaño*) -5,127*ln(*Temperatura*)

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	,901[a]	,811	,807	,53244	1,800

a. Variables predictoras: (Constante), LNTemperatura, LNTamaño

b. Variable dependiente: LNProteínas

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	105,791	2	52,895	186,583	,000[a]
	Residual	24,664	87	,283		
	Total	130,455	89			

a. Variables predictoras: (Constante), LNTemperatura, LNTamaño

b. Variable dependiente: LNProteínas

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados			Correlaciones			Estadísticos de colinealidad	
		B	Error típ.	Beta	t	Sig.	Orden cero	Parcial	Semiparcial	Tolerancia	FIV
1	(Constante)	11,876	2,317		5,125	,000					
	LNTamaño	2,649	,138	,916	19,226	,000	,858	,900	,896	,956	1,046
	LNTemperatura	-5,127	,876	-,279	-5,851	,000	-,088	-,531	-,273	,956	1,046

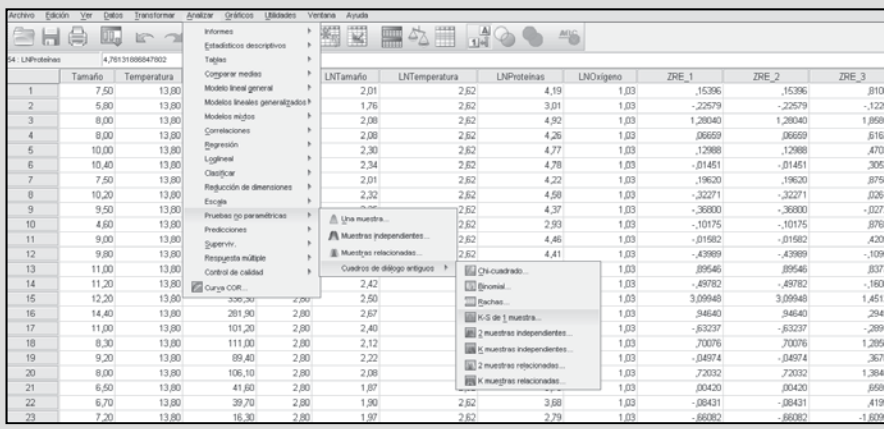
a. Variable dependiente: LNProteínas

A veces la opción por pasos no funciona bien y no escoge las variables que dan lugar al mejor modelo. Por ello, es conveniente realizar pruebas con diferentes combinaciones de las variables transformadas y no transformadas para obtener el mejor modelo usando la opción «Introducir» en el paso 10, y así incluir todas las variables independientes que se indiquen. En este caso, se hicieron todas las pruebas posibles y no se obtuvo un modelo mejor que el anterior.

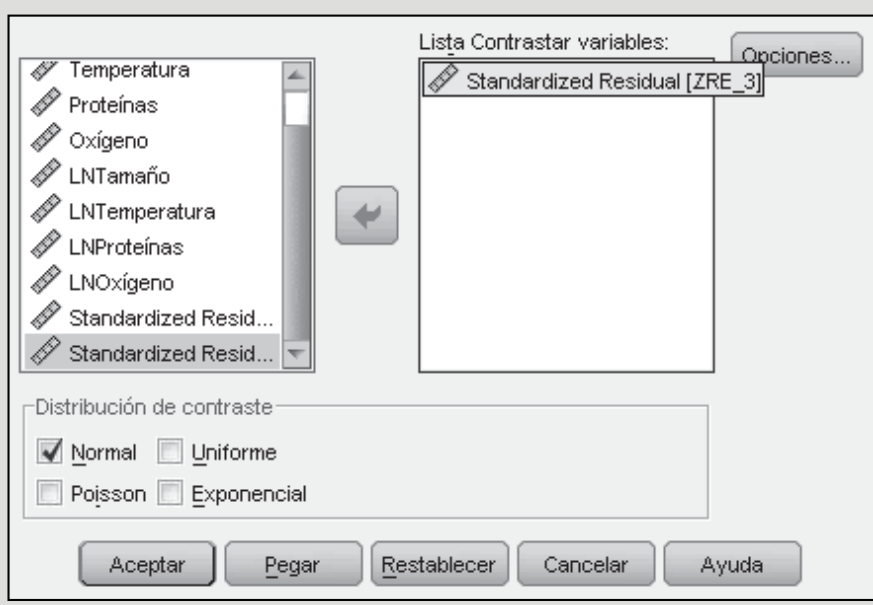
Hemos comprobado la colinealidad (véase paso 7), pero ahora debemos ver si es correcto usar el estadístico *F* y su nivel de significación comprobando los otros tres requisitos, es decir, que los residuos tengan distribución Normal, homocedasticidad y que no exista autocorrelación entre ellos.

CUADRO XI.2. *(Continuación)*[SPSS]

Paso 12. Distribución normal de los residuos. Volviendo a la página principal veremos que se han creado nuevas variables que contienen los residuos tipificados. Los residuos del tercer análisis que hemos realizado son «ZRE_3». Para ver la distribución Normal de los residuos entramos en «Analizar» ⇨ «Pruebas no paramétricas» ⇨ «KS de 1 muestra», para hacer el test de Kolmogorov-Smirnov.



Nos aparece la siguiente ventana en la cual, dentro de «Lista Contrastar variables:», introducimos la variable de la que queremos ver si tiene una distribución Normal, en este caso el último residuo, que corresponde al tercer análisis realizado «Standardized Residual(ZR_3)».



CUADRO XI.2. *(Continuación)*[SPSS]

Prueba de Kolmogorov-Smirnov para una muestra

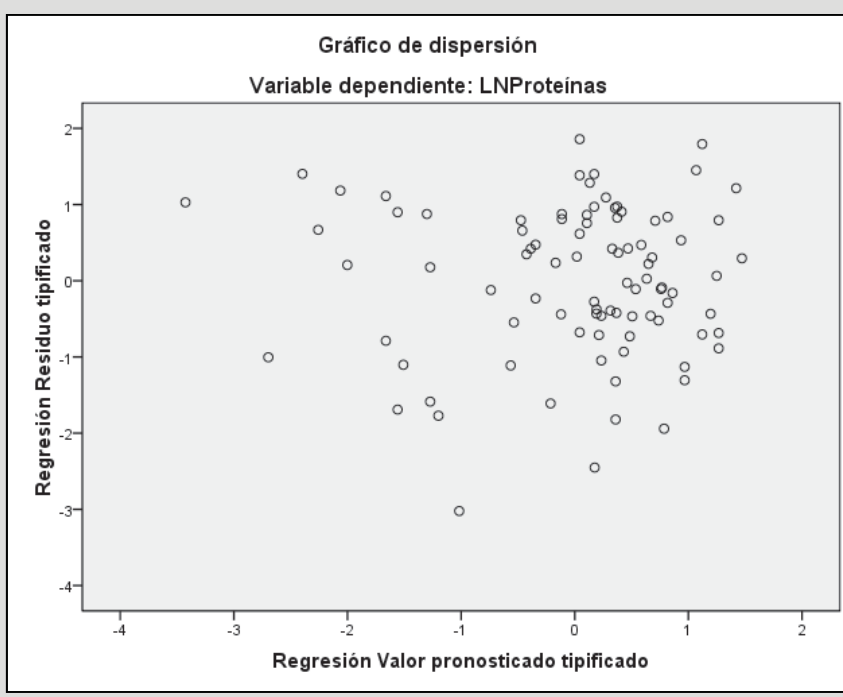
		Standardized Residual
N		90
Parámetros normales[a,b]	Media	,0000000
	Desviación típica	,98870020
Diferencias más extremas	Absoluta	,078
	Positiva	,049
	Negativa	-,078
Z de Kolmogorov-Smirnov		,738
Sig. asintót. (bilateral)		,647

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

Se observa que p = 0,647, por tanto, se acepta la hipótesis nula de que los residuos presentan una distribución normal.

Paso 13. Homocedasticidad de los residuos. Como se observa en la gráfica siguiente, la variabilidad de los residuos se mantiene más o menos constante a lo largo de todo el rango de los valores pronosticados tipificados. Por tanto, aceptamos que existe homocedasticidad en los residuos.



CUADRO XI.2. *(Continuación)*[SPSS]

Paso 14. Autocorrelación de los residuos. En el paso 11, en la tabla «Resumen del modelo», se ve el valor del contraste $\hat{d}$ = 1,8. Las tablas de Durbin-Watson (Tabla 7 del Apéndice I o archivo **Tabla 7.doc**) para n = 90 y α = 0,05 proporcionan los valores de d_L = 1,612 y d_U = 1,703 (considerando dos variables independientes).

- Si $0 < \hat{d} < d_L$ se rechaza H_0 y aceptamos la existencia de autocorrelación positiva.
- Si $d_L < \hat{d} < d_U$ el contraste no es concluyente.
- Si $d_U < \hat{d} < 4 - d_U$ se acepta H_0 y no hay autocorrelación.
- Si $4 - d_U < \hat{d} < 4 - d_L$ el contraste no es concluyente.
- Si $4 - d_L < \hat{d} < 4$ se rechaza H_0 y aceptamos la existencia de autocorrelación negativa.

Por lo tanto, como d_U (1,612) < $\hat{d}$ (1,8) < 4 - d_U $\hat{d}$ (2,297) se acepta H_0 y no hay autocorrelación entre los residuos.

XI.1.4. OTRAS REGRESIONES SIMPLES O MÚLTIPLES NO LINEALES

Muchas veces las variables se ajustan a un tipo de regresión que no es ninguna de las comunes que vienen descritas en la mayoría de los programas estadísticos. Sin embargo, algunos programas estadísticos permiten ver si la relación entre una variable dependiente y una o varias variables independientes se ajusta a una determinada ecuación. Para ello, es necesario tener una idea previa de la ecuación que mejor se puede ajustar a la tendencia que observamos entre las variables. En este apartado se describirán algunas de las relaciones no lineales que se observan más frecuentemente entre distintas variables.

XI.1.4.1. Curva logística

Muchos fenómenos en la naturaleza se ajustan a una curva logística como por ejemplo los cambios en la abundancia de una población en el tiempo (Smith & Smith, 2000), el avance científico a lo largo del tiempo dentro de una determinada línea de investigación (Solla Price, *Little Science, Big Science*, 1963 -citado en Callon y col., 1995), el ciclo de vida de uso que se le da a una línea de ferrocarril (Ingleda & Coto, 2003), etc. (Figura XI.3).

La curva logística viene definida por la siguiente ecuación:

$$Y = \frac{a}{1 + e^{(b - cX)}}$$

La constante *a* establece el límite superior de la curva y es igual a *K* o capacidad de carga de la población (número máximo de individuos de esa población) en el caso de que se modele a una población (Figura XI.3.). La constante *c* determina la pendiente, y en el caso de poblaciones es igual a *r* o tasa de crecimiento de la población. La constante *b* define el tamaño de la llamada fase de latencia en el caso de poblaciones, la fase con valores más pequeños antes de iniciar la pendiente hasta alcanzar el valor máximo.

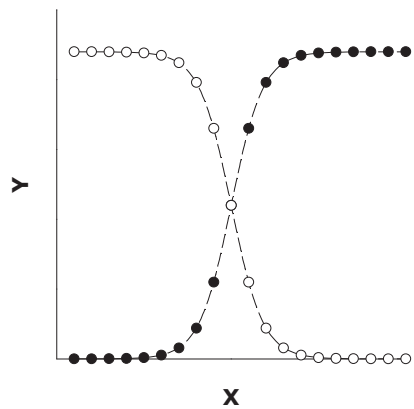


Figura XI.3. Curva logística con pendiente negativa (○) y positiva (●).

XI.1.4.2. Curva de crecimiento de von Bertalanffy

La curva de crecimiento de muchos organismos se ajusta bien a la ecuación de von Bertalanffy (Figura XI.4):

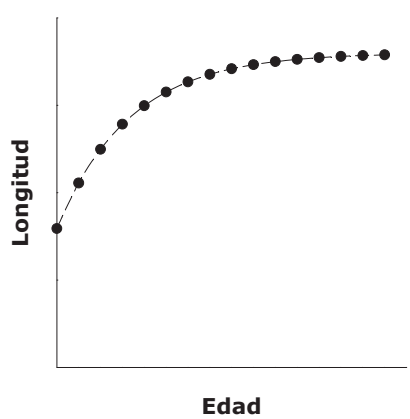


Figura XI.4. Curva de crecimiento de von Bertalanffy.

$$L_t = L_\infty \left(1 - \exp^{-k(t-t_0)}\right)$$

donde L_t es la longitud del individuo a la edad t, L_∞ es la longitud máxima que alcanza el individuo cuando el crecimiento cesa, k es la constante de crecimiento expresada en tiempo^{-1} y t_0 es la edad hipotética que tendría un individuo que tuviese un tamaño cero. Por lo tanto, la variable dependiente es L_t, la variable independiente la edad del individuo (t) y las constantes de esta ecuación son k, L_∞ y t_0.

XI.1.4.3. Curva de crecimiento de Gompertz

En algunas especies, como por ejemplo los peces, crustáceos y moluscos, en las primeras fases del ciclo de vida (fase larvaria) el crecimiento puede ajustarse mejor a la ecuación de Gompertz (Hernandez-Llamas & Ratkowsky, 2004) (Figura XI.5):

$$L_t = ae^{(-be^{(-ct)})}$$

donde L_t es la longitud del individuo a la edad t, a es una constante que representa el valor máximo asintótico de la curva, t es la edad del individuo y, b y c son las otras dos constantes de la ecuación.

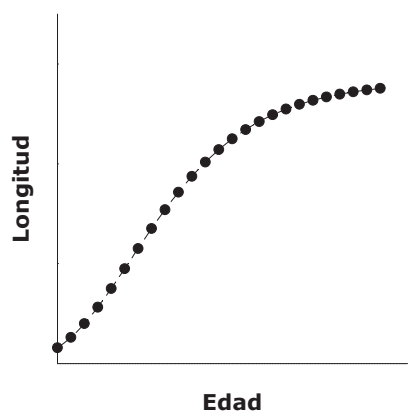


Figura XI.5. Curva de crecimiento de Gompertz.

XI.1.4.4. Relación entre tasas y variables

La tasa de crecimiento de una población bacteriana, de levaduras o de una fitoplanctónica a distintas concentraciones de nutrientes o de temperatura, la tasa fotosintética de plantas a diferentes intensidades de luz, entre otros muchos ejemplos, se ajustan bien a la ecuación de Monod (Figura XI.6):

$$\mu = \frac{\mu_{\max} P}{K_s + P}$$

donde μ es la tasa expresada en tiempo^{-1} y P es la variable (luz, temperatura, concentración de nutrientes, etc.) que provoca la variación de la tasa (de crecimiento, fotosintética, etc.).

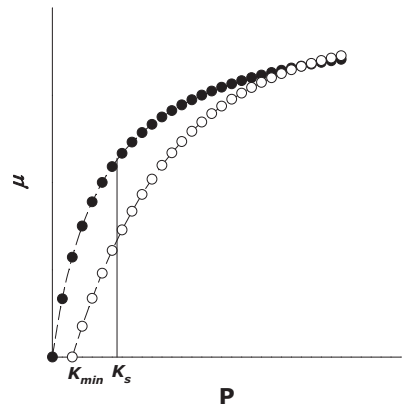


Figura XI.6. Ecuación de Monod (●) y ecuación que incluye K_{min} (○).

Las constantes de la ecuación son μ_{max} (en tiempo^{-1}) que es la tasa máxima asintótica de la ecuación y K_s que indica el valor de la variable P cuando se alcanza la mitad de la tasa máxima.

La ecuación de Monod presenta el problema de que la tasa solo alcanza el valor cero cuando el valor de la variable también es cero. Sin embargo, se observa frecuentemente que la tasa puede hacerse cero antes de que el valor de la variable explicativa sea cero. Por ejemplo, la tasa de crecimiento poblacional del fitoplancton se suele hacer cero a concentraciones de nutrientes (fosfato, nitrógeno o silicato) superiores a cero (Frangópulos y col. 2004). Esta concentración mínima de nutrientes a la que la tasa de crecimiento es cero, tiene gran importancia en términos de competencia entre las especies, porque aquellas especies para las que esta concentración es mayor significa que compiten peor por ese nutriente. La ecuación siguiente corrige ese problema de la ecuación de Monod ya que contempla la posibilidad de que la tasa se haga cero antes de que el parámetro llegue a cero (Figura XI.6):

$$\mu = \mu_{\max}\left(1 - e^{(-b(P-K_{\min}))}\right)$$

donde K_{min} es el valor del parámetro al cual la tasa es cero. Como se mencionó anteriormente, en los estudios de competencia, a medida que K_{min} es mayor, la especie es peor competidora por el recurso (la variable P de la ecuación).

Esta ecuación que incluye K_{min} admite que se incluyan tasas negativas, lo cual también es frecuente observarlo en la naturaleza. Por ejemplo, que el crecimiento de una población bacteriana sea negativo y, por tanto, se reduzca la abundancia de la población cuando la concentración de nutrientes sea muy baja. En el caso de la ecuación de Monod no existe la posibilidad de incluir tasas negativas.

En el Cuadro XI.3 se describe la forma de obtener cualquier tipo de regresión entre variables y si la relación es significativa.

CUADRO XI.3. Regresiones no lineales

Aplicación con R.

EJEMPLO. En el archivo **Cuadro XI.3.V.csv** hay datos de la tasa de crecimiento poblacional (días^{-1}) de tres especies de fitoplancton (*Alexandrium tamarense*, *A. andersoni* y *A. minutum*) a diferentes concentraciones de fosfato (μM PO_4^{3-}). El objetivo es determinar si es posible predecir el crecimiento de la población de la especie *A. tamarense* en base a la concentración de fosfato por medio de la siguiente ecuación:

$$\mu = \mu_{max}\left(1 - e^{(-b(Fosfato - K_{min}))}\right)$$

Después de invocar al programa R y seleccionar el directorio de trabajo como se explica en el Apéndice II, debemos abrir el *script* **Cuadro XI.3.R** que se muestra en la siguiente ventana.

```
# Cuadro XI.3
# Regresión no lineal

#########################################################
# Sección modificable por el usuario
#########################################################

#Lectura archivo de datos csv
datos<-read.csv2("Cuadro XI.3.V.csv",enc="latin1")

#Selección de variables

#Variable dependiente
varRespuesta<-"Tasa.Crecimiento"

#Variable independiente
varExplicativa<-"Conc.Fosfato"

#Fórmula del modelo. "valorex" es el nombre que debe llevar la vairable explicativa
formula<-paste("umax*(1-exp(-b*(valorex-Kmin)))")

#Valores iniciales de parámetros
#parametros<-list(umax=0.4,b=1,Kmin=0.05)
parametros<-NULL

#Variables de selección.
varSeleccion1<-list(c("Especie","A. tamarense"))
#varSeleccion1<-NULL

#varSeleccion2<-list(c("","",""))
varSeleccion2<-NULL

#Nombre del archivo en que se guardan los residuos
CSVr<-"Salida Cuadro XI.3.1.csv"

#Nombre del archivo en que se guardan los datos de la ecuación
CSVc<-"Salida Cuadro XI.3.2.csv"

#Nombre del archivo de salida
ArchivodeSalida<-"Salida Cuadro XI.3.txt"
```

CUADRO XI.3. *(Continuación)*[R]

Paso 1. En primer lugar es necesario importar los datos desde nuestro archivo csv.

```
datos<-read.csv2("Cuadro XI.3.V.csv",enc="latin1")
```

Paso 2. Se seleccionan las variables dependiente e independiente.

```
#Variable dependiente
varRespuesta<-"Tasa.Crecimiento"

#Variable independiente
varExplicativa<-"Conc.Fosfato"
```

Paso 3. Se especifica la función a calcular, en este caso «umax*(1-exp(-b*(valorex-Kmin)))»

```
formula<-paste("umax*(1-exp(-b*(valorex-Kmin)))")
```

Paso 4. Si se tiene una idea de los valores aproximados de los coeficientes de la regresión, lo más conveniente es indicarlos. En este caso no se especifica nada y se deja que el programa los estime.

```
#parametros<-list(umax=0.4,b=1,Kmin=0.05)
parametros<-NULL
```

Paso 5. También es posible seleccionar dos grupos en dos variables diferentes. En este ejemplo se hace la regresión sólo para la especie *A. tamarense*.

```
#Variables de selección.
varSeleccion1<-list(c("Especie","A. tamarense"))
#varSeleccion1<-NULL

#varSeleccion2<-list(c("","",""))
varSeleccion2<-NULL
```

Paso 6. Se especifica el nombre del archivo donde se guardarán los residuos (**Salida Cuadro XI.3.1.csv**), el archivo de los coeficientes del modelo de regresión (**Salida Cuadro XI.3.2.csv**) y, por último, el archivo con los cálculos realizados (**Salida Cuadro XI.3.3.txt**).

```
#Nombre del archivo en que se guardan los residuos
CSVr<-"Salida Cuadro XI.3.1.csv"
#Nombre del archivo en que se guardan los datos de la ecuación
CSVc<-"Salida Cuadro XI.3.2.csv"
#Nombre del archivo de salida
ArchivodeSalida<-"Salida Cuadro XI.3.txt"
```

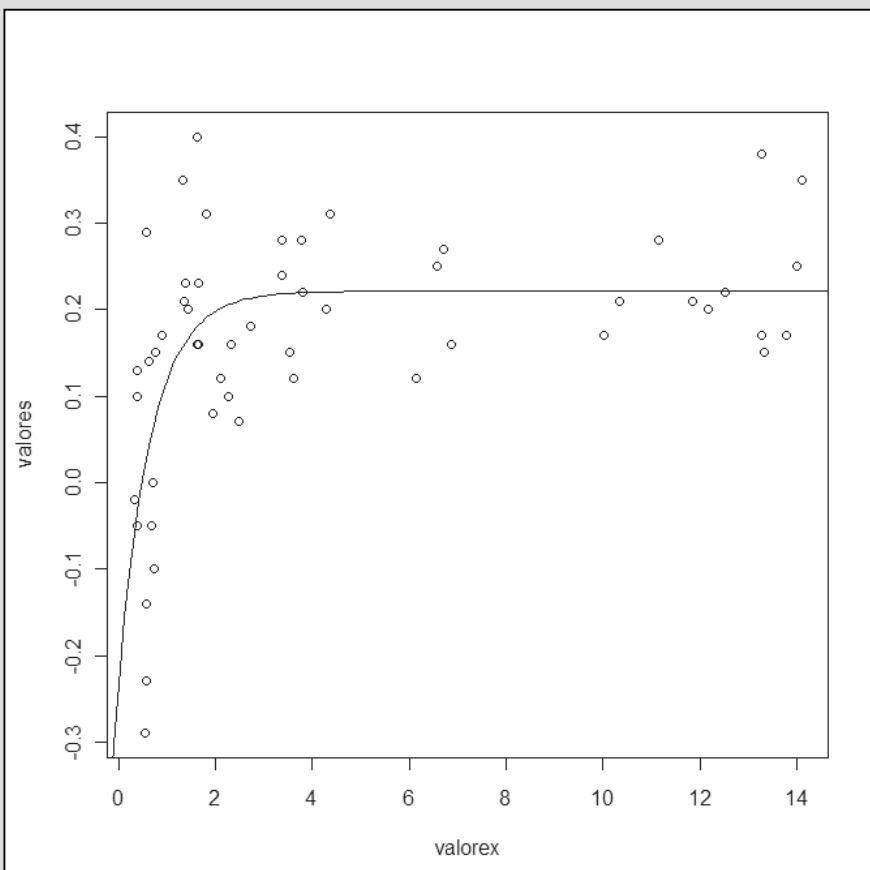
Paso 7. Por último, seleccionamos el contenido de este *script* pulsando *ctrl+A* y lo ejecutamos pulsando *ctrl+R*. Tal vez es necesario pulsar una vez con el ratón sobre la ventana que sale, para que aparezca la gráfica.

CUADRO XI.3. *(Continuación)*[R]

Paso 8. En el archivo **Salida Cuadro XI.3.txt** se observa que el r^2 = 0,379, y también aparecen los coeficientes del modelo.

```
Formula: valores ~ umax * (1 - exp(-b * (valorex - Kmin)))

Parameters:
     Estimate Std. Error t value Pr(>|t|)
umax  0.22110    0.02265   9.760 3.62e-13 ***
b     1.47932    0.68426   2.162   0.0354 *
Kmin  0.48027    0.10193   4.712 2.00e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1133 on 50 degrees of freedom

Number of iterations to convergence: 10
Achieved convergence tolerance: 6.834e-06

Valor.de.R^2:
    0.3796316
```

Paso 9. La gráfica muestra los datos y la curva obtenida.



CUADRO XI.3. *(Continuación)*[R]

Paso 10. En el archivo **Salida Cuadro XI.3.1.csv** se observan los residuos de la regresión, que permiten identificar los datos que peor se ajustan al modelo.

	x
1	-0,06870445
2	0,08961443
3	0,03396408
4	0,16917601
5	-0,01841175
6	-0,16459754
7	-0,05109679
8	-0,02109678
9	-0,05109663
10	-0,01952725
11	-0,13961968
12	0,04808431
13	-0,03316781
14	0,2596739
15	-0,09893979
16	0,06060559
17	0,19180525
18	0,13917601

Paso 11. En el archivo **Salida Cuadro XI.3.2.csv** se observan los coeficientes del modelo.

	x
umax	0,22109679
b	1,47932478
Kmin	0,48027312

Por lo tanto, el modelo de regresión no lineal entre la tasa de crecimiento (μ, d^{-1}) y la concentración de fosfato (μM PO_4^{3-}) para A. *tamarense* sería:

$$\mu = 0{,}221\left(1 - e^{(-1{,}479(Fosfato - 0{,}4802))}\right)$$

que indica una tasa de crecimiento máxima de 0,22 d^{-1} y que es necesario una concentración superior a 0,48 μM de fosfato para que el alga crezca.

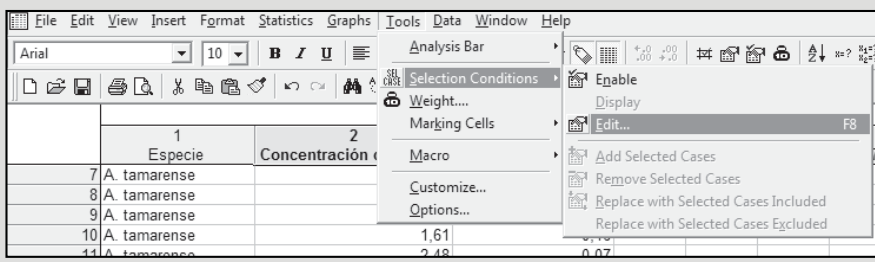
CUADRO XI.3. *(Continuación)* STATISTICA

Aplicación con STATISTICA

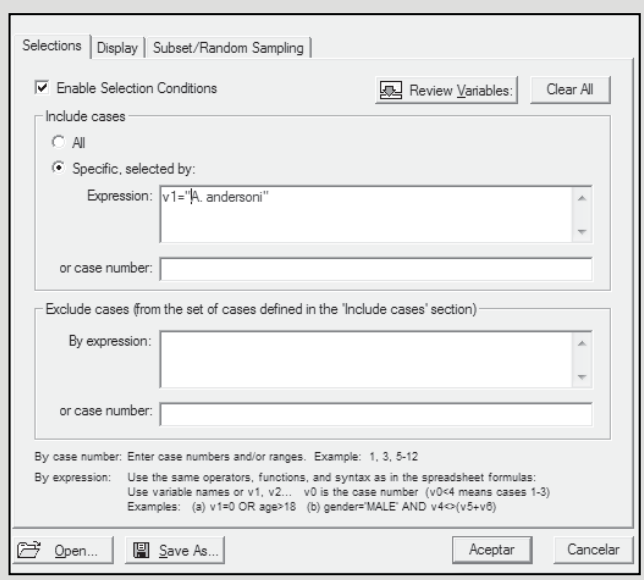
EJEMPLO. Se usa el mismo ejemplo que en la aplicación con R. En el archivo **Cuadro XI.3.sta** hay datos de la tasa de crecimiento poblacional (días^{-1}) de tres especies de fitoplancton (*A. tamarense*, *A. andersoni* y *A. minutum*) a diferentes concentraciones de fosfato (μM PO_4^{3-}). En este caso, el objetivo es determinar si es posible predecir el crecimiento de la población de la especie *A. andersoni* en base a la concentración de fosfato, por medio de la siguiente ecuación:

$$\mu = \mu_{max}\left(1 - e^{\left(-b(Fosfato - K_{min})\right)}\right)$$

Paso 1. Debido a que existen varias especies, es necesario primero definir la especie con la que se va a trabajar. Para seleccionar los casos vamos a «Tools» ⇨ «Selection Conditions» ⇨ «Edit…».

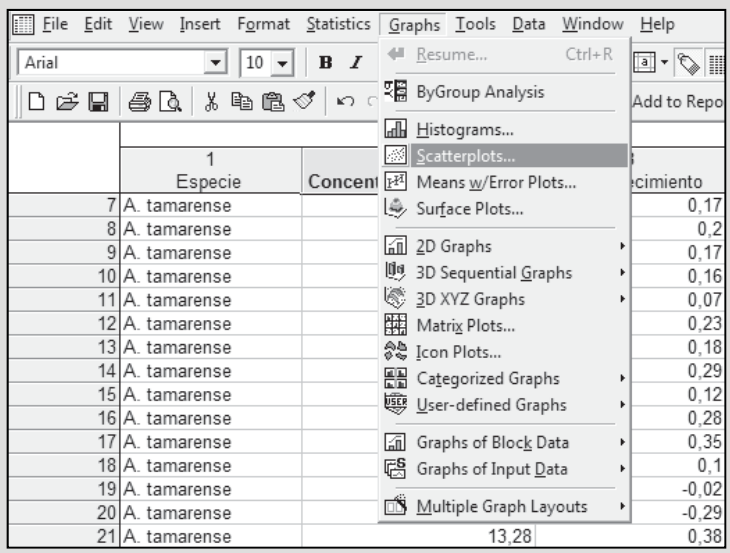


Paso 2. En la pestaña «Selections» se marca «Enable Selection Conditions» ⇨ «Specific, selected by:» donde se selecciona la variable o variables a analizar, usando OR y AND si se quieren combinar varias condiciones. En este caso se selecciona «v1="A. andersoni"».

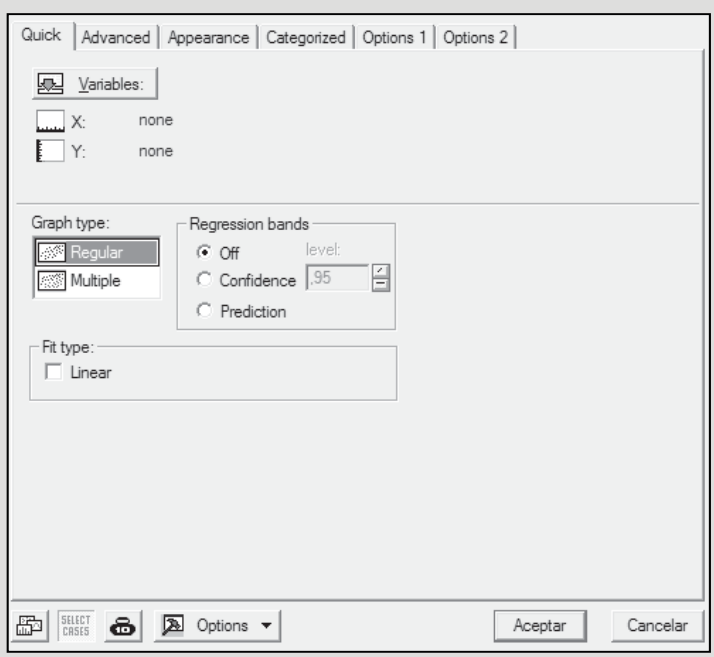

Selections Display Subset/Random Sampling
Enable Selection Conditions
Review Variables:
Clear All
Include cases
All
Specific, selected by:
Expression: v1="A. andersoni"
or case number:
Exclude cases (from the set of cases defined in the 'Include cases' section)
By expression:
or case number:
By case number: Enter case numbers and/or ranges. Example: 1, 3, 5-12
By expression: Use the same operators, functions, and syntax as in the spreadsheet formulas:
Use variable names or v1, v2... v0 is the case number (v0<4 means cases 1-3)
Examples: (a) v1=0 OR age>18 (b) gender='MALE' AND v4<>(v5+v6)
Open...
Save As...
Aceptar
Cancelar

CUADRO XI.3. *(Continuación)*[STATISTICA]

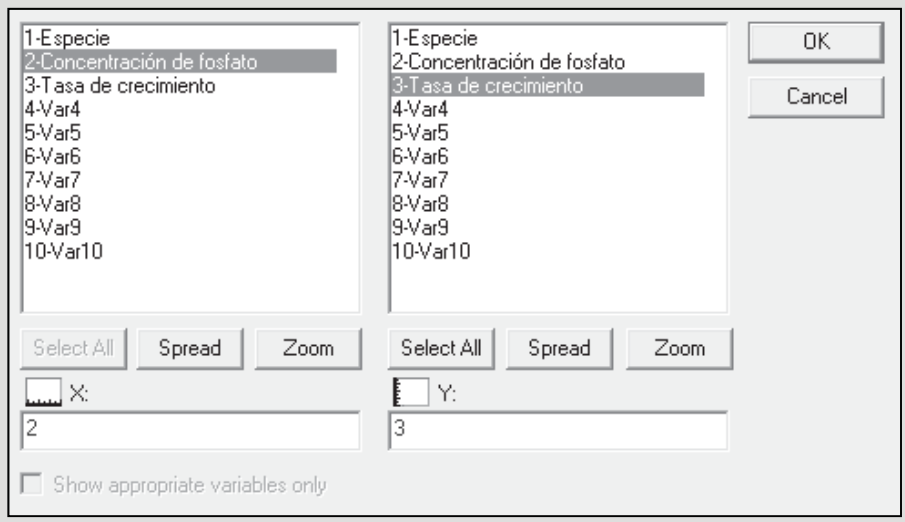
Paso 3. A continuación es importante graficar los resultados para ver el tipo de relación que existe entre las variables y si hay posibles *outliers*. Para ello vamos a «Graphs» ⇨ «Scatterplots…».



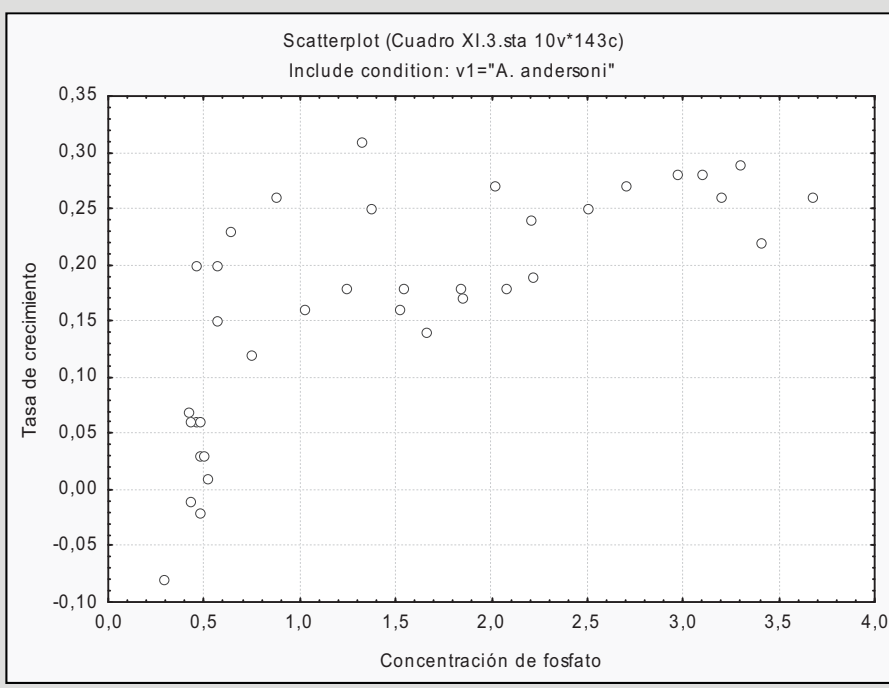
Paso 4. En la siguiente ventana, en «Fit type:» se desmarca «Linear» y al pulsar en «Variables» aparece la ventana que se muestra posteriormente donde seleccionamos las variables «Concentración de fosfato» y «Tasa de crecimiento».



CUADRO XI.3. *(Continuación)*[STATISTICA]

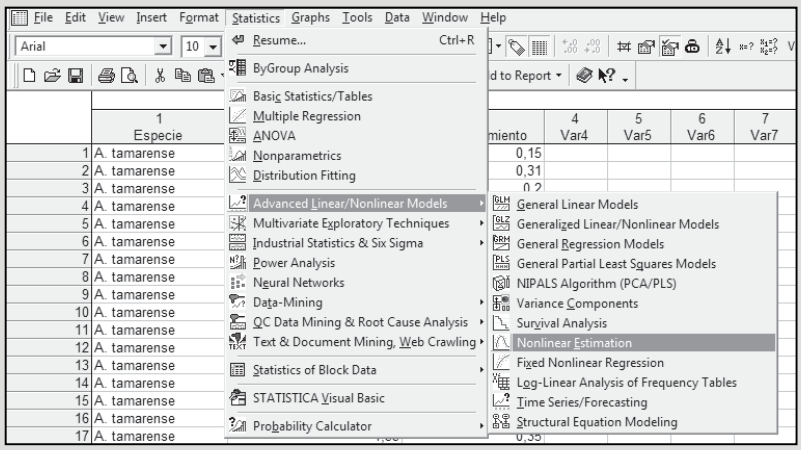


Paso 5. Se observa una relación entre ambas variables en la cual, a medida que aumenta la concentración de nutrientes la tasa de crecimiento es mayor, aunque parece que se alcanza un máximo, es decir, la tasa de crecimiento no aumenta más aunque siga aumentando la concentración de fosfato.

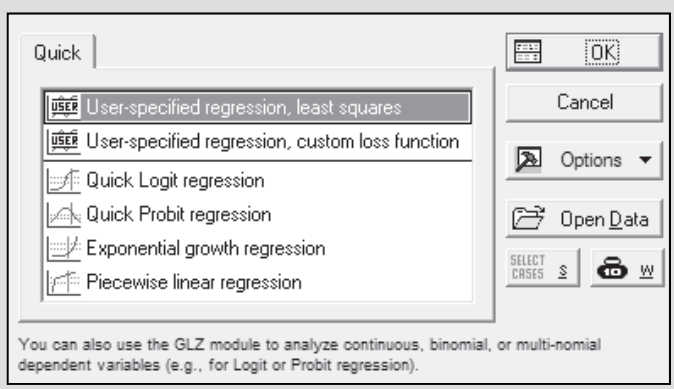


CUADRO XI.3. *(Continuación)* STATISTICA

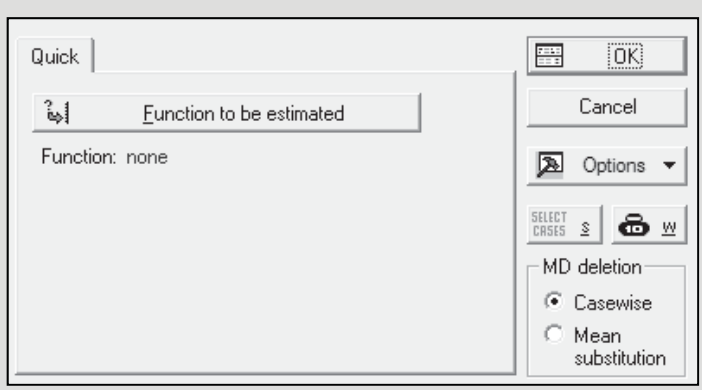
Paso 6. Para obtener la ecuación del modelo se va a «Statistics» ⇨ «Advanced Linear/Nonlinear Models» ⇨ «Nonlinear Estimation».



Paso 7. En la ventana siguiente se selecciona «User-specified regression, least squares».

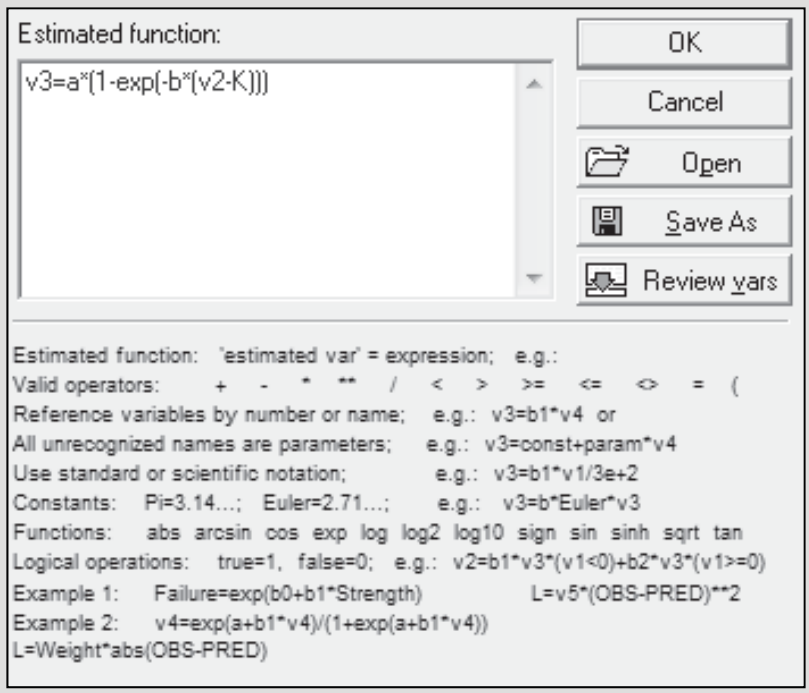


Paso 8. A continuación se selecciona «Function to be estimated».

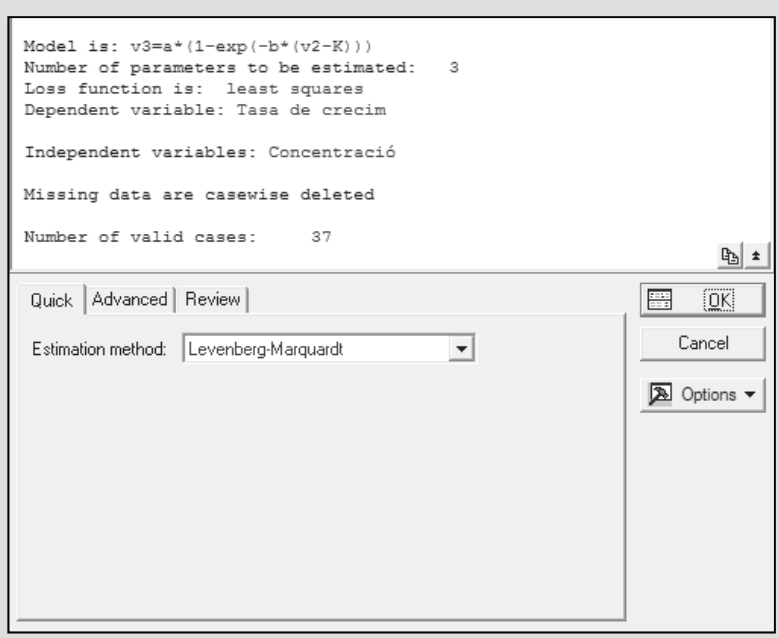


CUADRO XI.3. *(Continuación)*[STATISTICA]

Paso 9. Ponemos la ecuación del modelo que queremos estimar: «v3=a*(1-exp(-b*(v2-K)))».

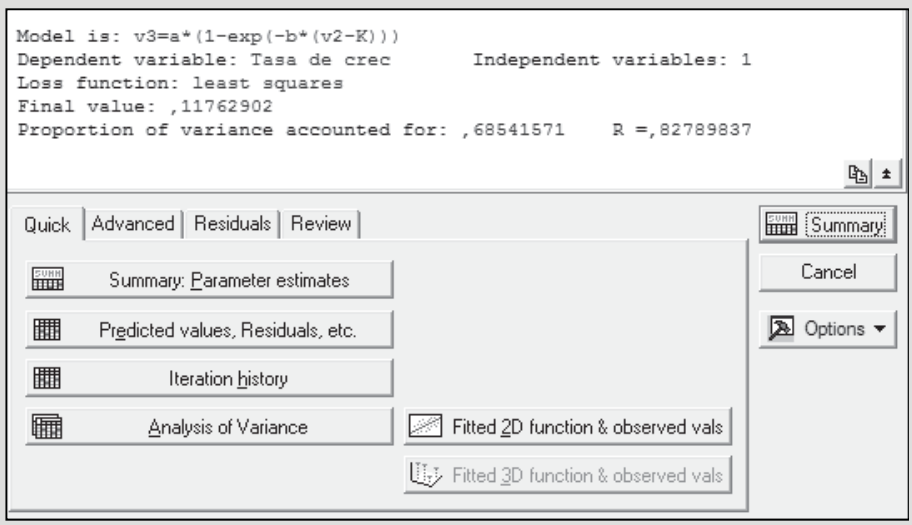

Estimated function:
v3=a*(1-exp(-b*(v2-K)))
OK
Cancel
Open
Save As
Review vars
Estimated function: 'estimated var' = expression; e.g.:
Valid operators: + - * ** / < > >= <= <> = (
Reference variables by number or name; e.g.: v3=b1*v4 or
All unrecognized names are parameters; e.g.: v3=const+param*v4
Use standard or scientific notation; e.g.: v3=b1*v1/3e+2
Constants: Pi=3.14...; Euler=2.71...; e.g.: v3=b*Euler*v3
Functions: abs arcsin cos exp log log2 log10 sign sin sinh sqrt tan
Logical operations: true=1, false=0; e.g.: v2=b1*v3*(v1<0)+b2*v3*(v1>=0)
Example 1: Failure=exp(b0+b1*Strength) L=v5*(OBS-PRED)**2
Example 2: v4=exp(a+b1*v4)/(1+exp(a+b1*v4))
L=Weight*abs(OBS-PRED)

Paso 10. Aparece la siguiente ventana y pulsamos «OK».



CUADRO XI.3. *(Continuación)*[STATISTICA]

Paso 11. Como r^2 = 0,68, la concentración de fosfato explica el 68% de la variación observada en la tasa de crecimiento diaria de *A. andersoni*. Como en los modelos no lineales no se cumplen en general las hipótesis necesarias para asegurar la distribución *F* del estadístico resultante del ANOVA, no tiene sentido usar el estadístico *F* ni la probabilidad, que nos indicaría si la relación entre la variable dependiente y las independientes es significativa.



Paso 12. Por último, en la pestaña «Quick» se pulsa en «Summary: Parameter estimates» y se obtienen los coeficientes de la ecuación:

	Model is: v3=a*(1-exp(-b*(v2-K))) (Cuadro XI.3.sta) Dep. Var. : Tasa de crecimiento Level of confidence: 95.0% (alpha=0.050) Include condition: v1="A. andersoni"					
	Estimate	Standard error	t-value df = 34	p-level	Lo. Conf Limit	Up. Conf Limit
a	0,230441	0,013664	16,86526	0,000000	0,202673	0,258209
b	3,710131	1,006554	3,68597	0,000788	1,664568	5,755695
K	0,375524	0,028393	13,22586	0,000000	0,317822	0,433226

Por lo tanto, el modelo de regresión no lineal entre la tasa de crecimiento (μ, d^{-1}) y la concentración de fosfato ($\mu M\ PO_4^{3-}$) sería:

$$\mu = 0{,}2304\left(1 - e^{(-3{,}71(Fosfato - 0{,}375))}\right)$$

que indica una tasa de crecimiento máxima de 0,23 d^{-1} y que es necesaria una concentración superior a 0,37 μM de fosfato para que el alga crezca.

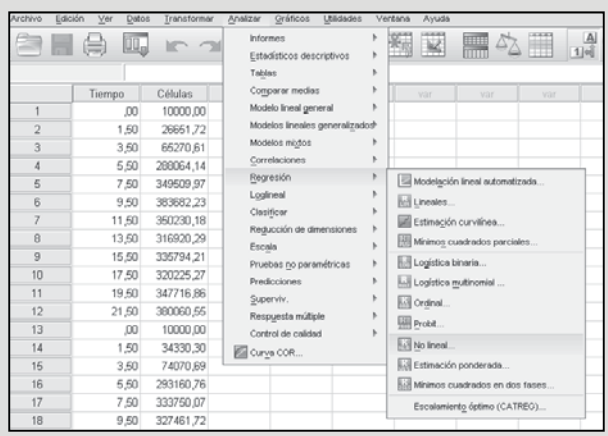
CUADRO XI.3. *(Continuación)*[SPSS]

Aplicación con SPSS.

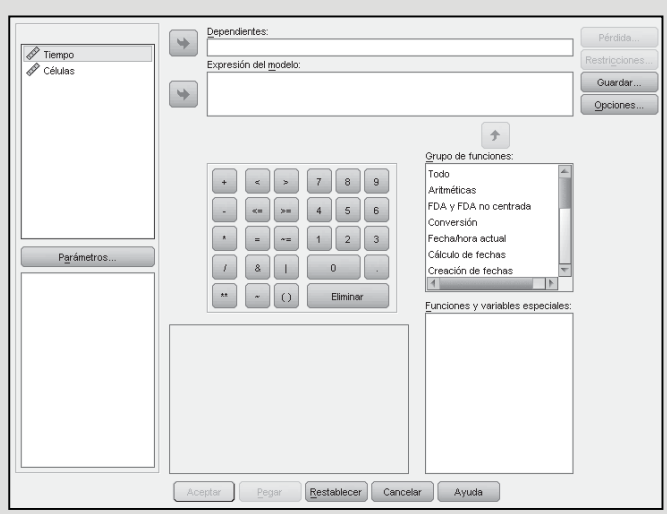
EJEMPLO. El crecimiento de una población de fitoplancton sigue normalmente una curva de crecimiento logístico. En el archivo **Cuadro XI.3.sav** se muestra la abundancia de fitoplancton (células ml^{-1}) en cultivo a lo largo del tiempo (en días), teniendo varias réplicas para cada tiempo. Se quiere estimar la curva logística de crecimiento, que viene definida por la siguiente ecuación:

$$Células = \frac{a}{1 + e^{(b - cTiempo)}}$$

Paso 1. Se selecciona «Analizar» ⇨ «Regresión» ⇨ «No lineal…».

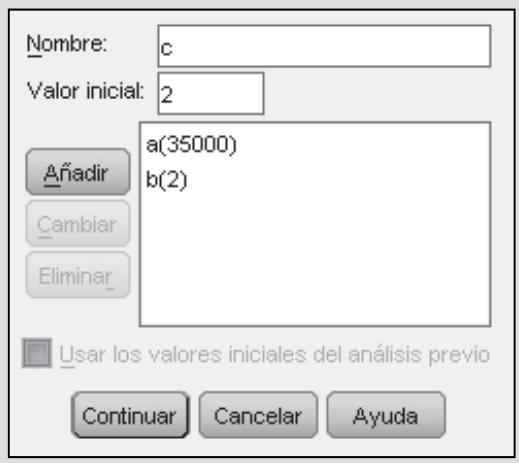


Paso 2. Nos aparecerá el siguiente cuadro en el cual es necesario, en primer lugar, definir los valores iniciales que le damos a las constantes de la ecuación y, para ello, entramos en «Parámetros».

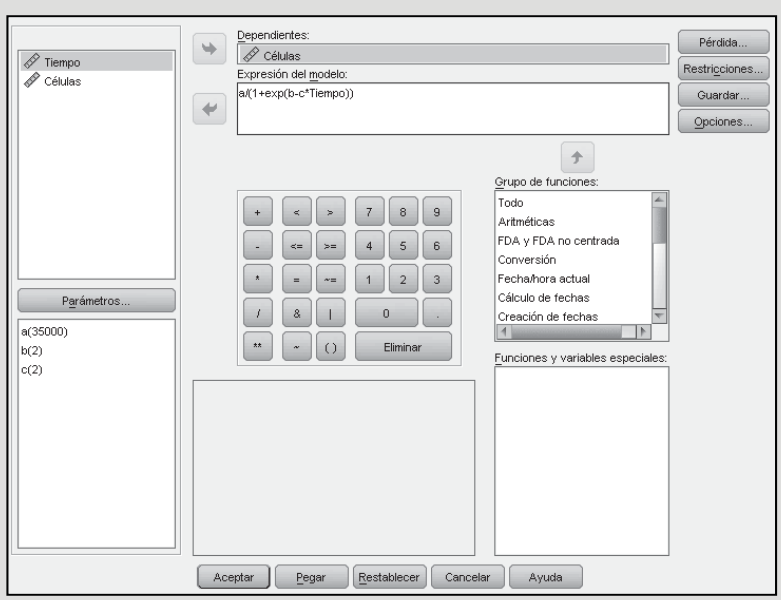


CUADRO XI.3. *(Continuación)*[SPSS]

Paso 3. Nos aparece el siguiente cuadro donde hay que introducir el nombre de la constante en «Nombre:» y el valor de la misma en «Valor inicial:». En nuestro caso será $a = 35000$, $b = 2$ y $c = 2$. Es importante introducir valores cercanos a lo que puede ser el resultado final, ya que si no se corre el riesgo de que el programa no sea capaz de ajustar la ecuación.



Paso 4. Una vez introducidas las constantes, el paso siguiente es introducir en «Expresión del modelo:» la ecuación que deseamos ajustar a nuestras variables, como se muestra en la siguiente ventana «a/(1+exp(b-c*Tiempo))». Es importante mencionar que es posible introducir varias variables independientes, es decir, es posible realizar una regresión múltiple.



CUADRO XI.3. *(Continuación)*[SPSS]

Paso 5. En los resultados nos aparece la tabla «Historial de iteraciones», con las iteraciones que el programa ha realizado hasta llegar al cálculo de los parámetros.

Historial de iteraciones[b]

Número de iteraciones[a]	Suma de cuadrados residual	Parámetro		
		a	b	c
1.0	3,272E12	35000,000	2,000	2,000
1.1	4,090E12	311381,134	4,202	-17,518
1.2	4,090E12	169232,867	5,713	-1,755
1.3	2,302E12	83751,041	2,579	1,475
2.0	2,302E12	83751,041	2,579	1,475
2.1	1,066E12	180861,396	3,347	,621
3.0	1,066E12	180861,396	3,347	,621
3.1	3,219E10	338598,181	4,304	1,085
4.0	3,219E10	338598,181	4,304	1,085
4.1	2,156E10	343521,411	5,211	1,175
5.0	2,156E10	343521,411	5,211	1,175
5.1	2,062E10	342437,403	5,847	1,323
6.0	2,062E10	342437,403	5,847	1,323
6.1	2,053E10	342318,247	6,075	1,374
7.0	2,053E10	342318,247	6,075	1,374
7.1	2,052E10	342262,118	6,134	1,387
8.0	2,052E10	342262,118	6,134	1,387
8.1	2,052E10	342247,327	6,148	1,390
9.0	2,052E10	342247,327	6,148	1,390
9.1	2,052E10	342243,932	6,151	1,390
10.0	2,052E10	342243,932	6,151	1,390
10.1	2,052E10	342243,179	6,152	1,390
11.0	2,052E10	342243,179	6,152	1,390
11.1	2,052E10	342243,013	6,152	1,390

Las derivadas se calculan numéricamente.

a. El número de iteraciones mayores se muestra a la izquierda del decimal, mientras que el número de iteraciones menores se encuentra a la derecha del decimal.

b. La ejecución se detuvo después de 24 evaluaciones de modelos y 11 evaluaciones de derivadas, ya que la reducción relativa entre sumas residuales sucesivas de cuadrados es, como mucho, SSCON = 1,00E-008.

CUADRO XI.3. *(Continuación)*[SPSS]

En la tabla «Estimaciones de los parámetros» se muestra el valor de las constantes que en este caso serían a = 342243, b = 6,152 y c = 1,39.

Estimaciones de los parámetros

Parámetro	Estimación	Error típico	Intervalo de confianza al 95%	
			Límite inferior	Límite superior
a	342243,013	3853,134	334482,403	350003,623
b	6,152	,608	4,927	7,376
c	1,390	,137	1,114	1,667

Por lo tanto la ecuación quedaría de la siguiente forma:

$$Abundancia = \frac{342243}{1 + e^{(6{,}152 - 1{,}39 * Tiempo)}}$$

En la parte inferior de la tabla «ANOVA» observamos el Coeficiente de Determinación «R cuadrado» (r^2), que es un buen indicador de la bondad del ajuste e indica la proporción de la varianza de la variable dependiente explicada por la regresión: en este caso su valor es 0,976 por lo que la regresión explica el 97,6% de la variabilidad, y tan solo el 2,4% restante es variabilidad residual no explicada por el ajuste. Un valor igual a 1 se obtendría únicamente cuando todos los puntos están en la curva, y por lo tanto el modelo explica exactamente los datos de la muestra, y en general se consideran satisfactorios porcentajes superiores a 0,90 ó 0,95.

La tabla de ANOVA contiene las sumas de cuadrados (de la regresión y residual) que permiten calcular el coeficiente r^2, así como los cuadrados medios. No se construye con ellos el cociente, estadístico que permitiría verificar la significación estadística de la regresión, ya que en general no se cumplen -en los modelos no lineales- las hipótesis necesarias para asegurar la distribución F del estadístico resultante. Por lo tanto, no se muestra el estadístico F ni la probabilidad, que nos indicaría si la relación entre la variable dependiente y las independientes es significativa.

ANOVA[a]

Origen	Suma de cuadrados	gl	Medias cuadráticas
Regresión	4,069E12	3	1,356E12
Residual	2,052E10	45	4,561E8
Total sin corrección	4,090E12	48	
Total corregido	8,423E11	47	

Variable dependiente: Células

a. R cuadrado = 1 - (Suma de cuadrados residual) / (Suma corregida de cuadrados) = ,976.

XI.2. Modelos de regresión para variables dependientes cualitativas

En el caso de que la variable dependiente sea cualitativa un modelo adecuado es el de regresión logística.

Se utiliza ampliamente en investigación clínica ya que permite estimar la probabilidad de ocurrencia de un proceso en función de ciertas variables, permitiendo evaluar la influencia de las variables independientes sobre la variable dependiente, dando como resultado una probabilidad.

La variable dependiente es siempre cualitativa pero las variables independientes pueden ser continuas, discretas, categóricas, dicotómicas o una mezcla de todas ellas.

Al igual que ocurría con la regresión para variables dependientes cuantitativas es necesario que no exista multicolinealidad entre las diferentes variables independientes y las observaciones de la muestra deben ser independientes entre sí. Sin embargo, no se requiere que los residuos presenten una distribución Normal ni la hipótesis de homocedasticidad (varianza constante de los residuos).

La variable dependiente puede ser dicotómica (0 si el hecho no ocurre y 1 si ocurre) o politómica (hay varias categorías), dando lugar a dos tipos diferentes de regresión logística, la binomial o multinomial, respectivamente.

XI.2.1. REGRESIÓN LOGÍSTICA BINOMIAL

La variable dependiente es:

$$Y_i = \log\left(\frac{\Pi_i}{1-\Pi i}\right)$$

donde Π_i es la probabilidad de que en el caso *i* se produzca el evento estudiado e $Y_i = \alpha + \beta X_i$ es el valor de la variable dependiente en el caso *i*.

Expresado en forma de regresión:

$$\Pi_i = \frac{1}{1+e^{-(\alpha+\beta X_i)}}$$

donde α y β son los coeficientes de la ecuación.

En caso de existir varias variables predictoras (independientes) la regresión se transforma en:

$$\Pi_i = \frac{1}{1+e^{\left[-(\alpha+\sum_{j=1}^{k}\beta_j X_{ji})\right]}}$$

donde Π_i es la probabilidad de que se produzca el evento estudiado en el caso *i*, *k* es el número de variables predictoras, α es un coeficiente, β_j es el coeficiente de la variable predictora *j* y X_{ji} es el valor de la variable predictora *j* en el caso *i*.

En caso de que Π_i sea mayor de 0,5 se asume (a efectos de predicción) que el evento se produce y si es menor que 0,5 que no se produce.

CUADRO XI.4. Regresión logística binomial

Aplicación con R.

EJEMPLO. Una empresa desea saber si es posible predecir si una persona compra un producto, en base a su estrato económico y su edad. Los datos están en el archivo **Cuadro XI.4.V.csv**.

Después de invocar al programa R y seleccionar el directorio de trabajo como se explica en el Apéndice II, debemos abrir el *script* **Cuadro XI.4.R** que se muestra en la siguiente ventana.

```
# Cuadro XI.4
# Regresión logística binomial
#########################################################
# Sección modificable por el usuario
#########################################################

#Cargar archivo de datos
datos<-read.csv2("Cuadro XI.4.V.csv",header=TRUE,enc="latin1")

#Seleccionar variables respuesta y predictoras.
#Las predictoras deben escribirse entre las comillas en el vector "varpred".
varesp<-"Producto"
varpred<-paste("Estrato.económico","Edad",sep="+")

#Seleccionar dentro de la variable respuesta el valor
#respecto al que hay que transformar la variable respuesta en binaria (0/1)
gruporesp<-("Compra")

#Seleccionar si se quiere trabajar con todas las
#variables independientes, o utilizar métodos "forward" y "backwards".
variables<-c("todas")
variables<-c("backwards","forward")

#Variables de selección de datos.
#varSeleccion1<-list(c("Sexo","Hombre"))
varSeleccion1<-NULL
#varSeleccion2<-list(c("","",""))
varSeleccion2<-NULL

#Nombre del archivo de salida
ArchivodeSalida<-"Salida Cuadro XI.4.txt"

#Nombre del archivo csv con los valores de los coeficientes
CSV<-"Salida Cuadro XI.4-1.csv"

#Nombre del archivo csv con los valores de probabilidad
CSVp<-"Salida Cuadro XI.4-2.csv"
```

Paso 1. En primer lugar es necesario importar los datos desde nuestro archivo csv.

datos<-read.csv2("Cuadro XI.4.V.csv",header=T,encoding="latin1")

CUADRO XI.4. *(Continuación)*[R]

Paso 2. Se selecciona la variable dependiente, en este caso «Producto», y las independientes, que son «Estrato.económico» y «Edad». Para que R reconozca que la variable independiente es categórica (en este caso «Estrato.económico»), es necesario que las clases estén identificadas al menos con una letra.

```
varesp<-"Producto"
varpred<-paste("Estrato.económico","Edad",sep="+")
```

Paso 3. A continuación hay que especificar cual es la clase que será referencia en la variable dependiente. En este caso se especifica que sea la clase «Compra», es decir, las personas que compran el producto.

```
gruporesp<-("Compra")
```

Paso 4. El modelo se puede realizar incluyendo todas las variables independientes que se especifican en el paso 2, o elegir la opción «backwards» y «forward», que escoge el mejor modelo y sólo incluye las variables que son significativas. En cualquier caso, siempre es conveniente realizar el análisis incluyendo todas las variables independientes, para comprobar que el método por pasos elige el modelo más predictivo. En primer lugar escogemos la opción «backwards» y «forward».

```
#variables<-c("todas")
variables<-c("backwards","forward")
```

Paso 5. También es posible seleccionar dos grupos en dos variables diferentes. En este ejemplo se hace el modelo de regresión para hombres y mujeres conjuntamente.

```
#varSeleccion1<-list(c("Sexo","Hombre"))
varSeleccion1<-NULL
#varSeleccion2<-list(c("","",""))
varSeleccion2<-NULL
```

Paso 6. Se especifica el nombre del archivo donde se guardarán los cálculos realizados, el cual es un archivo de texto, también el archivo donde se guardarán los coeficientes del modelo y el archivo donde se guardan las probabilidades de los casos, que tienen formato csv y se pueden abrir con Excel.

```
ArchivodeSalida<-"Salida Cuadro XI.4.txt"
CSV<-"Salida Cuadro XI.4-1.csv"
CSVp<-"Salida Cuadro XI.4-2.csv"
```

Paso 7. Por último, seleccionamos el contenido de este *script* pulsando *ctrl+A* y lo ejecutamos pulsando *ctrl+R*.

CUADRO XI.4. *(Continuación)*[R]

Paso 8. Se observa que el modelo es capaz de predecir correctamente el 97,87% de los casos de las personas que compran el producto y el 95,65% de los casos de las personas que no lo compran. En total acierta en el 96,77% de los casos y, por tanto, el modelo es bastante predictivo.

```
[1] "Porcentaje correctamente clasificado total"
[2] "96.7741935483871"
[1] "Porcentaje correctamente clasificado Producto Compra"
[2] "97.8723404255319"
[1] "Porcentaje correctamente clasificado Producto No Compra"
[2] "95.6521739130435"
```

Paso 9. Los coeficientes de la regresión, que se pueden ver en los archivos **Salida Cuadro XI.4.txt** y **Salida Cuadro XI.4.csv**, muestran que los compradores son personas jóvenes, ya que el coeficiente de la «Edad» es negativo, y con un estrato económico alto, ya que es positivo para el estrato de más de 3000 € y negativo para el estrato entre 1500 y 3000 €, aunque este último no es significativo (p = 0,914).

```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        20.1663     6.2016   3.252  0.00115 **
Estrato.económicoMás 3000 euros     6.0237     2.1965   2.742  0.00610 **
Estrato.económicoMenos de 1500 euros -0.2375     2.2219  -0.107  0.91487
Edad                               -0.6544     0.2027  -3.229  0.00124 **
```

	x
(Intercept)	20,1662931
Estrato.económicoMás 3000 euros	6,02365668
Estrato.económicoMenos de 1500 euros	-0,23751666
Edad	-0,65440011

El modelo vendría definido por la siguiente ecuación:

$$\Pi = \frac{1}{1 + e^{-(20{,}1663 - 0{,}6544 * Edad + 6{,}0237 * Estrato(>3000) - 0{,}2375 * Estrato(<1500)}}$$

Por ejemplo, la probabilidad de que una persona de 25 años y de estrato entre 1500 € y 3000 € compre el producto sería:

$$\Pi = \frac{1}{1 + e^{-(20{,}1663 - 0{,}6544 * 25 + 6{,}0237 * 0 - 0{,}2375 * 0)}} = 0{,}97$$

Como la probabilidad es mayor que 0,5 entonces el modelo predice que esa persona si compraría el producto.

Otro ejemplo, la probabilidad de que una persona de 44 años y de estrato > 3000 € compre el producto sería:

$$\Pi = \frac{1}{1 + e^{-(20{,}1663 - 0{,}6544 * 44 + 6{,}0237 * 1 - 0{,}2375 * 0)}} = 0{,}068$$

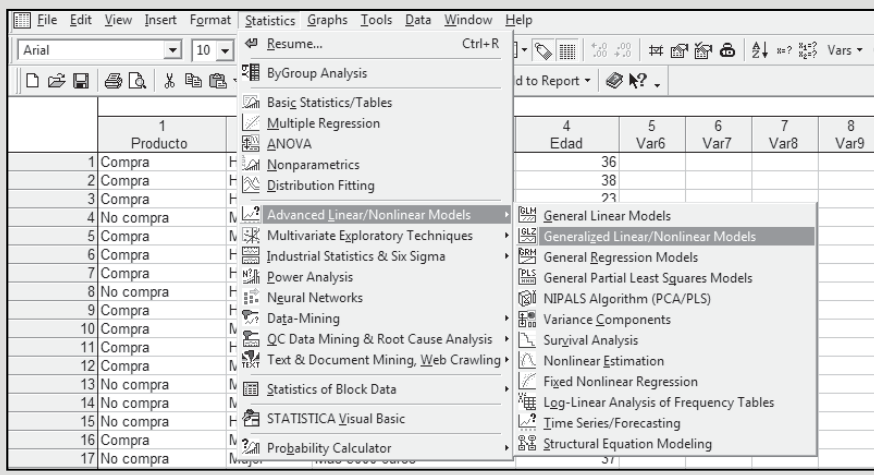
Como la probabilidad es menor que 0,5, entonces el modelo predice que esa persona no compraría el producto.

CUADRO XI.4. *(Continuación)*[STATISTICA]

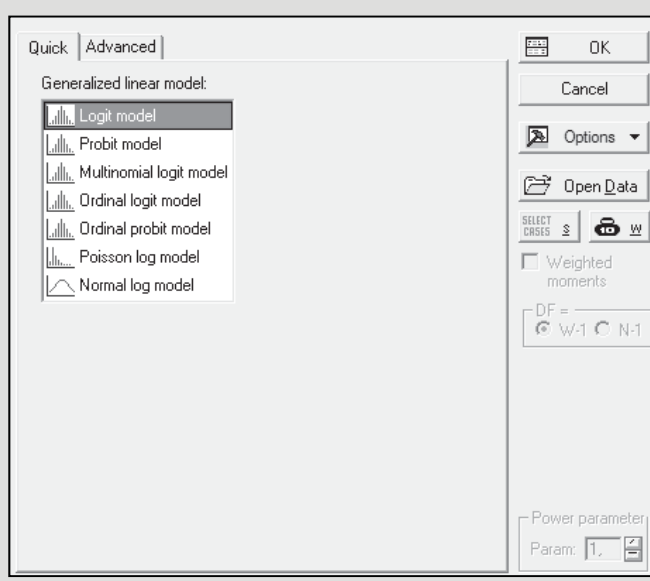
Aplicación con STATISTICA.

EJEMPLO. Vamos a utilizar el mismo ejemplo que en la aplicación con R. Una empresa desea saber si es posible predecir si una persona compra un producto, en base a su estrato económico y su edad. En este caso se hace el análisis considerando tanto a los hombres como a las mujeres. Los datos están en el archivo **Cuadro XI.4.sta**.

Paso 1. Se selecciona «Statistics» ⇨ «Advanced Linear/Nonlinear Models» ⇨ «Generalized Linear/Nonlinear Models».

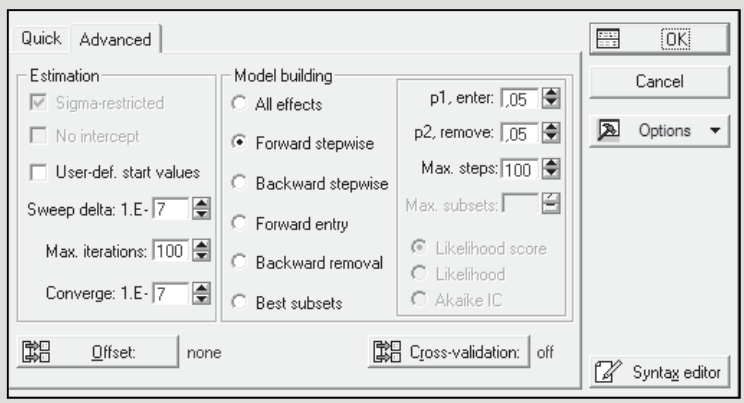


Paso 2. En la pestaña «Quick» se selecciona «Logit model».

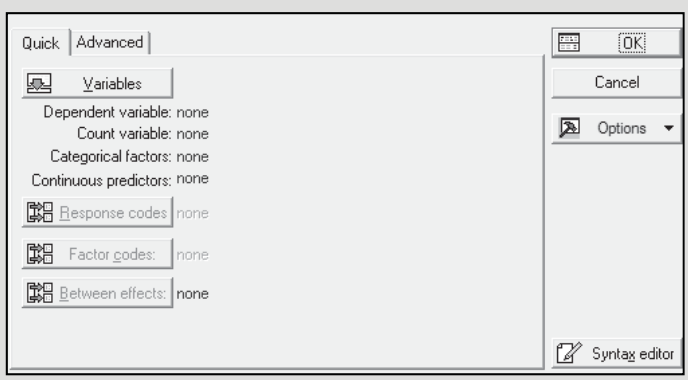


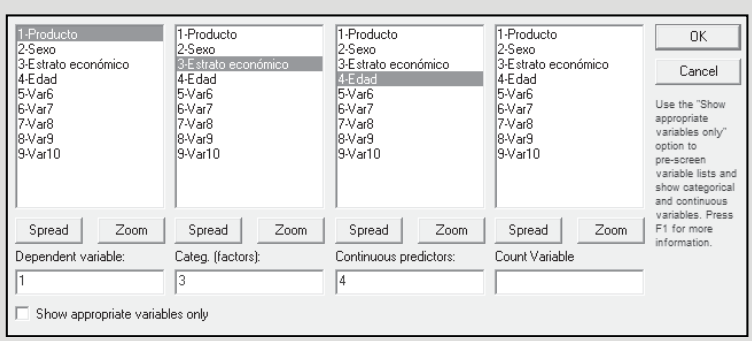
CUADRO XI.4. *(Continuación)* STATISTICA

Paso 3. En la pestaña «Advanced» se elige «Forward stepwise» para que incluya solo las variables independientes significativas. Las otras formas de construir el modelo se explicaron en el paso 7 de Cuadro XI.2. En cualquier caso, siempre es conveniente realizar el análisis incluyendo todas las variables independientes «All effects», para comprobar que el método por pasos elige el modelo más predictivo.



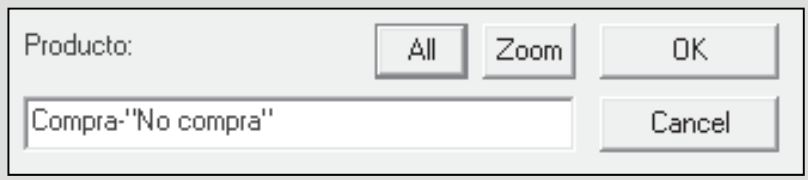
Paso 4. En la pestaña «Quick» se pulsa «Variables» y en la ventana que aparece se indica la variable dependiente, en este caso «Producto», «Estrato económico» es una variable categórica y «Edad» es continua.



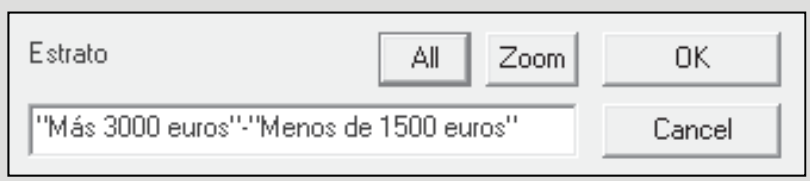


CUADRO XI.4. *(Continuación)*[STATISTICA]

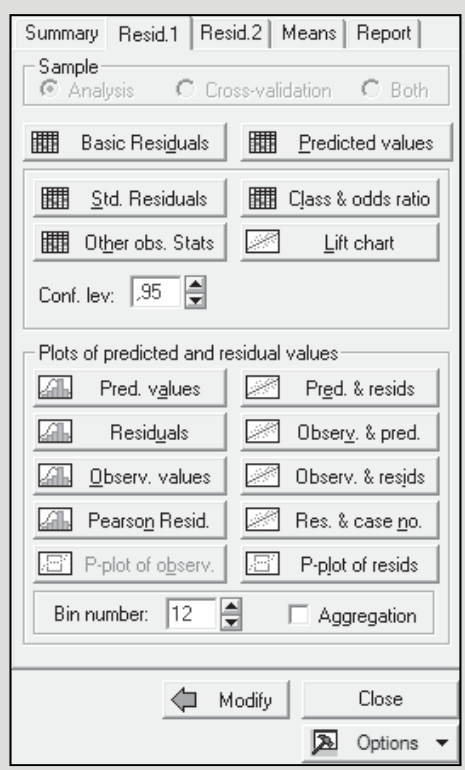
Paso 5. Para especificar las etiquetas de la variable dependiente, en la ventana del paso 4 se pulsa en «Response codes» y en la ventana que aparece se pulsa «All».



Para especificar las etiquetas de las variables independientes, en la ventana del paso 4 se pulsa en «Factor codes» y en la ventana que aparece se pulsa también «All».



Paso 6. En la pestaña «Resid.1» se pulsa «Class & odds ratio».

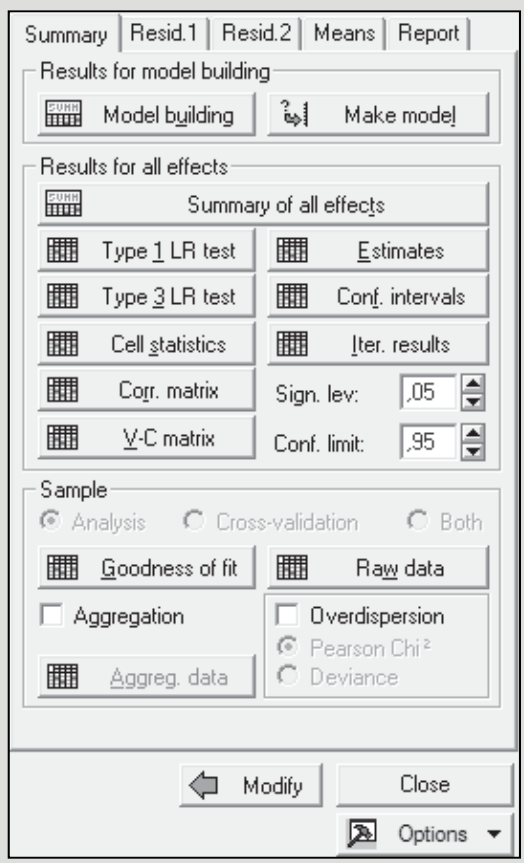


CUADRO XI.4. *(Continuación)* *STATISTICA*

Paso 7. Se observa que el modelo es capaz de predecir correctamente el 97,87% de los casos de las personas que compran el producto (sólo se equivoca en una persona) y el 95,65% de los casos de las personas que no lo compran (sólo predice mal dos personas). Por lo tanto, el modelo es bastante predictivo.

	Classification of cases (Cuadro XI.4.sta) Odds ratio: 1012,000000 Log odds ratio: 6,919684		
Observed	Predicted Compra	Predicted No compra	Percent correct
Compra	46	1	97,87234
No compra	2	44	95,65217

Paso 8. En la pestaña «Summary» se pulsa «Model building».



Paso 9. Se observa que en el primer paso incluye la variable «Edad» y en el segundo paso la variable «Estrato económico». Por tanto, ambas variables son significativas.

CUADRO XI.4. *(Continuación)* STATISTICA

Step	Producto - Model building results (Cuadro XI.4.sta) Distribution : BINOMIAL Link function: LOGIT Effect	Degr. of Freedom	Wald Stat.	Wald p	Score Stat.	Score p	Var. Status
Step 1	Edad	1			56,45486	0,000000	Entered
	Estrato económico	2			20,04724	0,000044	Out
Step 2	Edad	1	18,57811	0,000016			In
	Estrato económico	2			15,99723	0,000336	Entered
Step 3	Edad	1	10,42618	0,001242			In
	Estrato económico	2	7,91427	0,019118			In

Paso 10. En la ventana del paso 8 se pulsa «Summary of all effects». Para cada variable en el modelo se muestra el estadístico de Wald, que es el cuadrado del cociente entre el valor estimado del coeficiente y su desviación típica, y que se utiliza para contrastar la hipótesis de que el coeficiente correspondiente es nulo, ya que en ese supuesto sigue una distribución Chi-cuadrado. Si el valor *p* es menor que el nivel de significación elegido (por ejemplo 0,05), debemos conservar esa variable en el modelo. En este ejemplo vemos que la constante, la edad y el estrato económico son significativas.

Effect	Producto - Test of all effects (Cuadro XI.4.sta) Distribution : BINOMIAL Link function: LOGIT Degr. of Freedom	Wald Stat.	p
Intercept	1	10,63274	0,001111
Edad	1	10,42618	0,001242
Estrato económico	2	7,91427	0,019118

Paso 11. En la ventana del paso 8 se pulsa «Estimates». Se obtienen los coeficientes de la regresión, los cuales muestran que los compradores son personas jóvenes, ya que el coeficiente de la «Edad» es negativo, y con un estrato económico alto, ya que es positivo para el estrato de más de 3000 € y negativo para el estrato entre 1500 y 3000 €, aunque este último no es significativo.

Effect	Producto - Parameter estimates (Cuadro XI.4.sta) Distribution : BINOMIAL Link function: LOGIT Level of Effect	Column	Estimate	Standard Error	Wald Stat.	p
Intercept		1	22,09501	6,775972	10,63274	0,001111
Edad		2	-0,65440	0,202666	10,42618	0,001242
Estrato económico	Más 3000 euros	3	4,09494	1,513232	7,32293	0,006808
Estrato económico	Entre 1500 y 3000 euros	4	-1,92871	1,134008	2,89270	0,088982
Scale			1,00000	0,000000		

El modelo vendría definido por la siguiente ecuación:

$$\Pi = \frac{1}{1 + e^{-(22{,}095 - 0{,}6544 * Edad + 4{,}0949 * Estrato(>3000) - 1{,}9287 * Estrato(1500-3000)}}$$

CUADRO XI.4. *(Continuación)* *STATISTICA*

Por ejemplo, supongamos el primer caso de la tabla de datos, la probabilidad de que una persona de 36 años y de estrato > 3000 € compre el producto sería:

$$\Pi = \frac{1}{1 + e^{-(22{,}095 - 0{,}6544*36 + 4{,}0949*1 - 1{,}9287*0)}} = 0{,}93$$

Como la probabilidad es mayor que 0,5 entonces el modelo predice que esa persona si compraría el producto.

Otro ejemplo, la probabilidad de que una persona de 34 años y de estrato < 1500 € compre el producto sería:

$$\Pi = \frac{1}{1 + e^{-(22{,}095 - 0{,}6544*34 + 4{,}0949*0 - 1{,}9287*0)}} = 0{,}46$$

Como la probabilidad es menor que 0,5, entonces el modelo predice que esa persona no compraría el producto.

Paso 12. En la ventana del paso 6 se pulsa «Predicted values» y se obtienen las probabilidades que predice el modelo para cada uno de los casos y el intervalo de confianza al 95%.

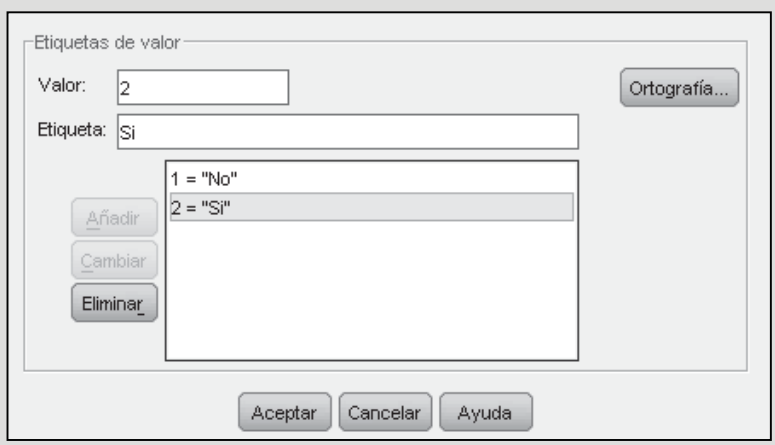
	Producto - Predicted Values (Cuadro XI.4.sta) Distribution : BINOMIAL Link function: LOGIT					
Case	Response '1' = Compra	Pred. Value	LINEAR Pred.	Standard Error	Lower CL 95, %	Upper CL 95, %
1	1,000000	0,932864	2,6315	1,08408	0,624053	0,991476
2	1,000000	0,789638	1,3227	0,91959	0,382341	0,957917
3	1,000000	0,994030	5,1151	1,74104	0,845890	0,999802
4	0,000000	0,060784	-2,7377	1,34101	0,004651	0,472680
5	1,000000	0,999801	8,5211	2,59319	0,968889	0,999999
6	1,000000	0,999617	7,8667	2,40404	0,959102	0,999997
7	1,000000	0,999985	11,1387	3,36554	0,989465	1,000000
8	0,000000	0,766507	1,1887	0,96752	0,330124	0,956270
9	1,000000	0,999896	9,1755	2,78423	0,976306	1,000000
10	1,000000	0,999801	8,5211	2,59319	0,968889	0,999999
11	1,000000	0,978253	3,8063	1,41572	0,737214	0,998616
12	1,000000	0,999996	12,4475	3,75777	0,993834	1,000000
13	0,000000	0,124635	-1,9493	1,20800	0,013166	0,603098
14	0,000000	0,000015	-11,1109	3,75539	0,000000	0,022968
15	0,000000	0,000005	-12,1368	4,28435	0,000000	0,023211
16	1,000000	0,110730	-2,0833	1,20384	0,011626	0,568607
17	0,000000	0,878377	1,9771	0,98457	0,511856	0,980293
18	0,000000	0,000093	-9,2817	3,13162	0,000000	0,041334
19	0,000000	0,470008	-0,1201	0,94600	0,121935	0,849925
20	1,000000	0,932864	2,6315	1,08408	0,624053	0,991476

CUADRO XI.4. *(Continuación)*[SPSS]

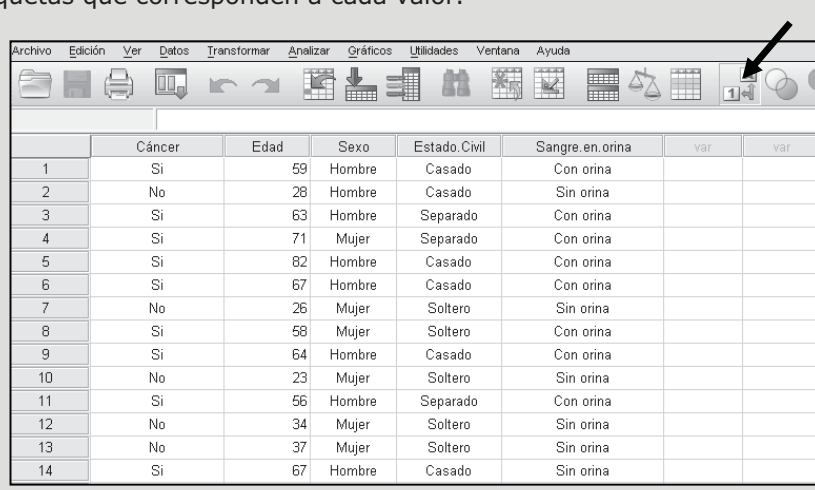
Aplicación con SPSS.

EJEMPLO. En el archivo **Cuadro XI.4.sav** hay datos de pacientes a los que se les ha diagnosticado un tipo de cáncer, así como de personas que no lo tienen. El objetivo es determinar si variables como el sexo, edad, estado civil (soltero, casado y separado) y presencia/ausencia de sangre en la orina, pueden ser indicadoras de la presencia del cáncer.

Paso 1. Al seleccionar la pestaña «Vista de variables» aparece la siguiente ventana donde es posible especificar el nombre de cada variable en «Nombre», el tipo de variable en «Tipo», etc. En «Valores» aparece la ventana que se muestra a continuación, donde es posible etiquetar a qué corresponde cada valor en cada una de las variables. Se muestra el ejemplo para la variable «Cáncer».



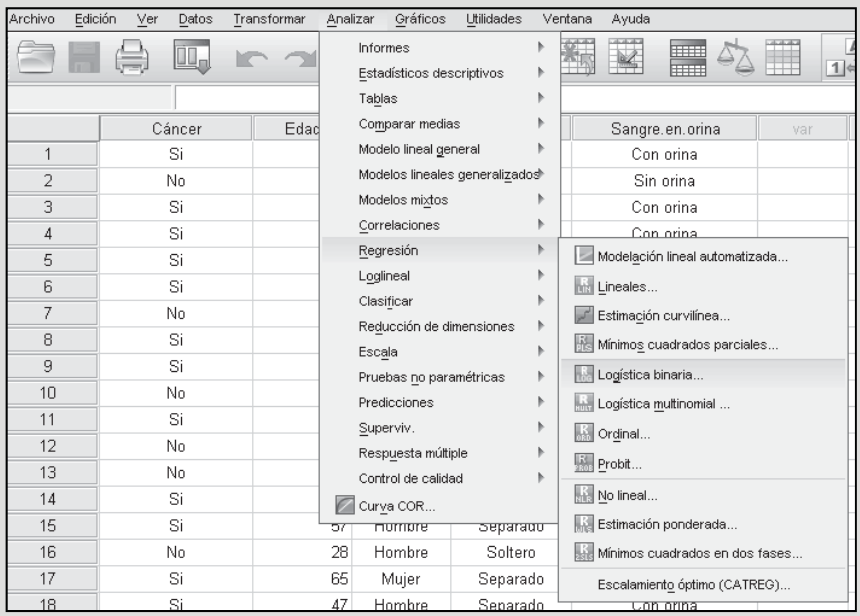
Paso 2. Al volver a la ventana de la pestaña «Vista de datos», si pulsamos donde indica la flecha aparecen, en vez de los números, las etiquetas que corresponden a cada valor.



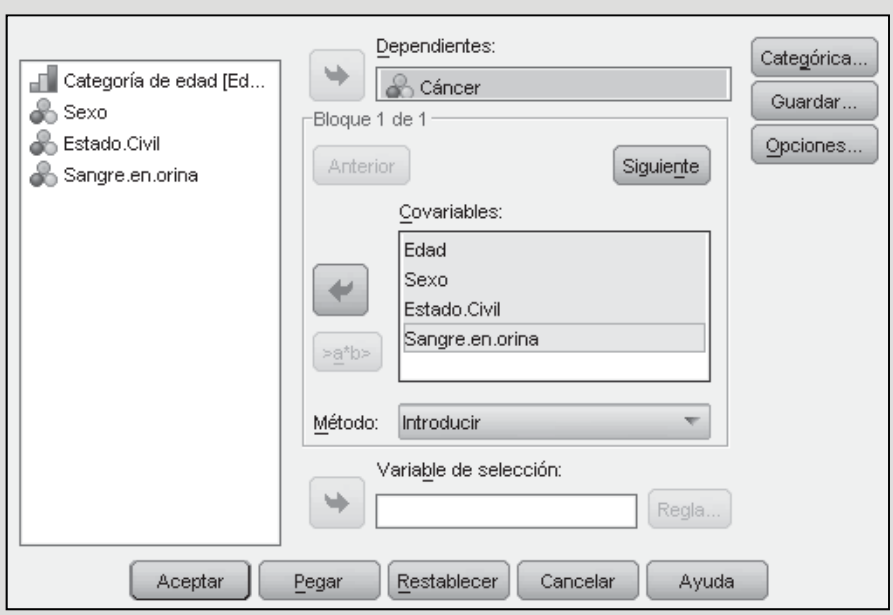
	Cáncer	Edad	Sexo	Estado.Civil	Sangre.en.orina	var	var
1	Si	59	Hombre	Casado	Con orina		
2	No	28	Hombre	Casado	Sin orina		
3	Si	63	Hombre	Separado	Con orina		
4	Si	71	Mujer	Separado	Con orina		
5	Si	82	Hombre	Casado	Con orina		
6	Si	67	Hombre	Casado	Con orina		
7	No	26	Mujer	Soltero	Sin orina		
8	Si	58	Mujer	Soltero	Con orina		
9	Si	64	Hombre	Casado	Con orina		
10	No	23	Mujer	Soltero	Sin orina		
11	Si	56	Hombre	Separado	Con orina		
12	No	34	Mujer	Soltero	Sin orina		
13	No	37	Mujer	Soltero	Sin orina		
14	Si	67	Hombre	Casado	Sin orina		

CUADRO XI.4. *(Continuación)*[SPSS]

Paso 3. Se selecciona «Analizar» ⇨ «Regresión» ⇨ «Logística binaria…».

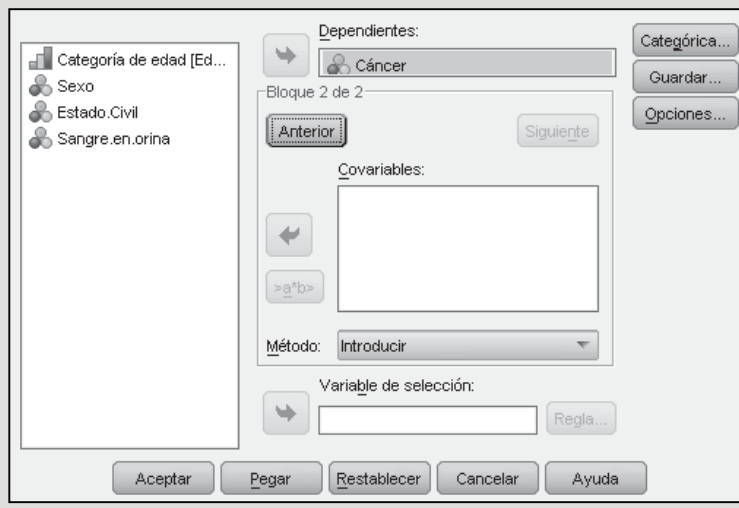


Paso 4. Nos aparecerá el siguiente cuadro en el cual es necesario, en primer lugar, definir la variable «Dependiente:» y las «Covariables:» (variables predictoras).



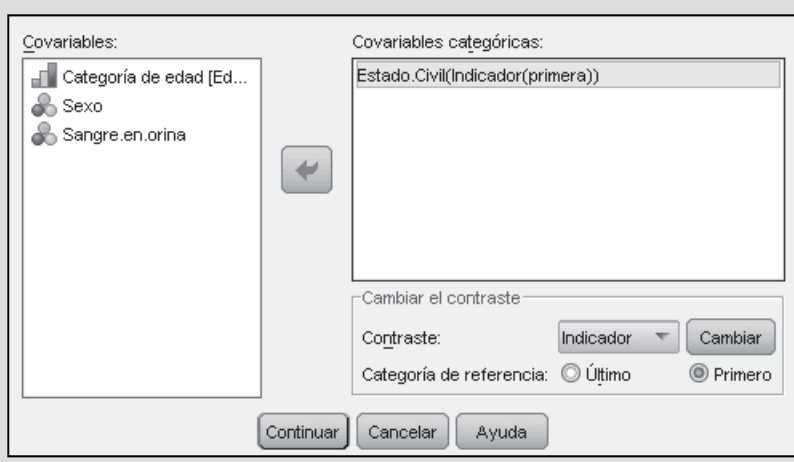
CUADRO XI.4. *(Continuación)*[SPSS]

En esta pantalla cabe destacar que es posible realizar distintos modelos de regresión (que se denominan bloques) al mismo tiempo, bien cambiando las variables independientes o el método de elección de variables. Para ello solamente debemos pulsar en el icono «Siguiente» y nos aparecerá la ventana que se muestra a continuación donde se pueden poner otras covariables diferentes y, así, realizar otra regresión.



Como podemos observar en la pantalla aparece el número de bloque indicando que es el Bloque 2 de 2, es decir, vamos a ejecutar 2 modelos y en esta pantalla se definirá el segundo. Podríamos añadir todos los bloques que quisiésemos, pero con nuestros datos vamos a trabajar sólo con el primer bloque, por lo que pulsamos sobre el icono «Anterior» para volver a la pantalla del Bloque 1.

Paso 5. Como algunas de nuestras variables independientes son categóricas pulsamos sobre el icono «Categórica...» para indicarlo.



CUADRO XI.4. *(Continuación)*[SPSS]

Las variables categóricas que son dicotómicas, «Sexo» (Hombre/Mujer) y «Sangre.en.orina» (Con sangre/Sin sangre), no es necesario considerarlas como categóricas. Por lo tanto, sólo pasamos al cuadro de «Covariables categóricas:» «Estado.Civil» (Soltero/Casado/Separado). Una vez hecho esto, aparece un paréntesis después de la variable en el que se nos indica cómo queremos seleccionar la categoría de referencia a la hora de realizar las comparaciones.

Existen diferentes contrastes que se pueden utilizar. Para cambiarlos pulsamos sobre «Cambiar». Las posibilidades son:

-*Indicador.* Presencia o ausencia de cada categoría.

-*Simple.* Cada categoría del predictor, excepto la propia categoría de referencia, se compara con la categoría de referencia.

-*Diferencia.* Cada categoría del predictor, excepto la primera categoría, se compara con el efecto promedio de las categorías anteriores.

-*Helmert.* Cada categoría del predictor, excepto la última categoría, se compara con el efecto promedio de las categorías subsiguientes.

-*Repetidas.* Cada categoría del predictor, excepto la primera categoría, se compara con la categoría que la precede.

-*Polinómico.* Contrastes polinómicos ortogonales. Se supone que las categorías están espaciadas equidistantemente. Los contrastes polinómicos sólo están disponibles para variables numéricas.

-*Desviación.* Cada categoría del predictor, excepto la categoría de referencia, se compara con el efecto global.

Para los contrastes «Indicador», «Simple» y «Desviación» es posible indicar si se desea que la categoría de referencia sea la primera o última de los datos. En nuestro caso, dejaremos «Indicador» y ponemos que tenga como referencia la primera categoría «Soltero». Para el cambio pulsar sobre el icono «Cambiar».

Paso 6. En la ventana del paso 4, pulsando en «Método:», se indica cómo queremos que se introduzcan las variables independientes en el modelo. Los métodos de introducción posibles son:

- *Introducir.* Procedimiento para la selección de variables en el que todas las variables de un bloque se introducen en un solo paso. Incluye todas las variables aunque no sean significativas.

- *Adelante:Condicional.* Contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación en base a la probabilidad de un estadístico de la razón de verosimilitud, que se basa en estimaciones condicionales de los parámetros.

CUADRO XI.4. *(Continuación)*[SPSS]

- *Adelante:RV.* Contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación en base al estadistico de la razón de verosimilitud, que se basa en estimaciones de la máxima verosimilitud parcial.

- *Adelante:Wald.* Método de selección por pasos hacia adelante que contrasta la entrada basándose en la significación del estadístico de puntuación y contrasta la eliminación basándose en la probabilidad del estadístico de Wald. El estadístico de Wald permite una prueba Chi-cuadrado para contrastar la hipótesis nula de que el coeficiente de cada variable independiente es cero.

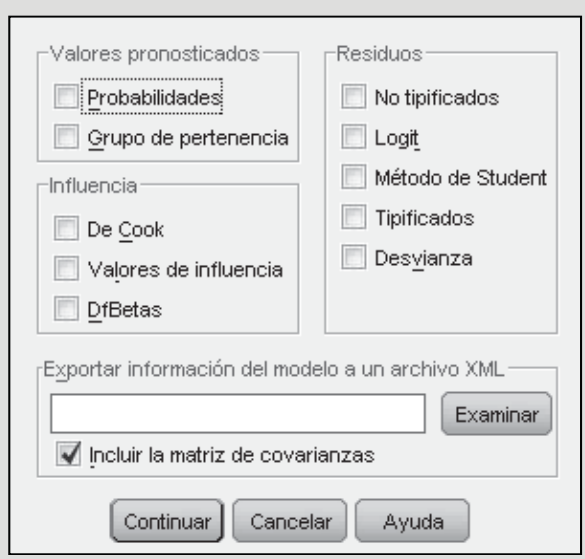
- *Atrás:Condicional.* Selección por pasos hacia atrás. El contraste para la eliminación se basa en la probabilidad del estadístico de la razón de verosimilitud, el cuál se basa a su vez en las estimaciones condicionales de los parámetros.

- *Atrás:RV.* Selección hacia atrás por pasos. El contraste para la eliminación se fundamenta en la probabilidad del estadístico de la razón de verosimilitud, el cual se basa en estimaciones de máxima verosimilitud parcial.

- *Atrás:Wald.* Selección por pasos hacia atrás. El contraste para la eliminación se basa en la probabilidad del estadístico de Wald.

En nuestro caso, seleccionaremos «Adelante:Condicional» ya que queremos que el modelo incluya solamente las variables significativas.

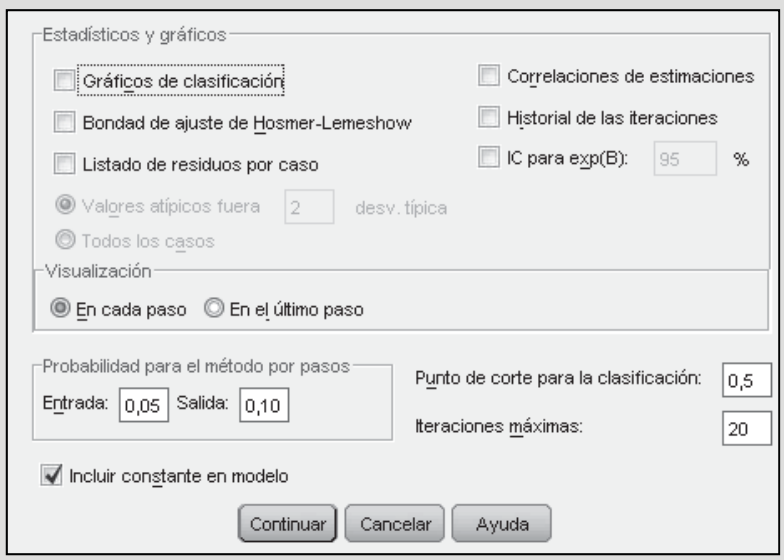
Paso 7. En la pantalla del paso 4, pulsando sobre el icono «Guardar» aparecen las posibles opciones para guardar los resultados de la regresión logística como nuevas variables.



CUADRO XI.4. *(Continuación)*[SPSS]

Es posible guardar la probabilidad de que se produzca el evento o el grupo de pertenencia -según el modelo- de cada caso. Además es posible solicitar la influencia de cada caso particular en los valores predichos: «De Cook» mide como cambian los residuos si excluimos cada caso, «Valores de influencia» la importancia de cada dato en el ajuste del modelo y «DfBetas» mide el cambio en el coeficiente de regresión excluyendo cada caso. También es posible obtener diferentes mediciones de los residuos.

Paso 8. Por último, pulsamos sobre el icono "Opciones" que nos permite personalizar algunas características del procedimiento.



En la sección «Estadísticos y gráficos» podemos indicar si queremos ver el «Gráfico de clasificación», así como la «Bondad de ajuste Hosmer-Lemeshow» (se basa en agrupar los casos en deciles de riesgo y comparar la probabilidad observada con la esperada dentro de cada decil). También podemos solicitar el «Listado de residuos por caso», las «Correlaciones de estimaciones», el «Historial de iteraciones» o el intervalo de confianza para los distintos β «IC para exp(B) ».

En la sección «Visualización» podemos indicar si queremos ver los resultados «En cada paso» o sólo «En el último paso».

También podemos indicar la probabilidad de «Entrada» y «Salida» de las covariables en los modelos por pasos, así como el «Nº máximo de iteraciones» y el «Punto de corte para la clasificación», es decir, a partir de qué probabilidad vamos a suponer o predecir que se produce el evento.

Con nuestros datos vamos a dejar las opciones que el programa trae por defecto.

CUADRO XI.4. *(Continuación)*[SPSS]

Paso 9. Una vez ya en los resultados, la tabla que aparece a continuación «Pruebas omnibus sobre los coeficientes del modelo», nos muestra el valor de los contrastes Chi- cuadrado entre nuestros datos y el modelo que hemos generado. Como nuestra regresión logística era por pasos, nos va a indicar los valores de Chi-cuadrado en los distintos pasos. El Chi-cuadrado de bloque no nos interesa, ya que en nuestro caso tenemos un único bloque, pero en caso de tener varios, nos indicaría si existen diferencias significativas entre los distintos bloques.

Pruebas omnibus sobre los coeficientes del modelo

				dimension2		
				Chi cuadrado	gl	Sig.
dimension0	Paso 1	dimension1	Paso	340,067	1	,000
			Bloque	340,067	1	,000
			Modelo	340,067	1	,000
	Paso 2	dimension1	Paso	160,486	1	,000
			Bloque	500,554	2	,000
			Modelo	500,554	2	,000
	Paso 3	dimension1	Paso	37,428	2	,000
			Bloque	537,982	4	,000
			Modelo	537,982	4	,000

En el primer paso (introducción de las cuatro variables independientes) observamos que el Chi-cuadrado del modelo es significativo con $p < 0{,}001$, es decir, las variables independientes describen la variable dependiente de forma significativa. En este caso el Chi-cuadrado de paso (comparación entre los pasos sucesivos de la regresión) es igual al del modelo, ya que lo compara con el modelo sin considerar variables.

En el paso 2 observamos que hay una reducción en el estadístico Chi-cuadrado entre pasos, y que hay diferencias significativas entre el modelo del paso 1 y el modelo del paso 2 con una variable más ($p < 0{,}001$) y que el modelo sigue siendo significativo ($p < 0{,}001$).

Ya en el último, el cambio en el valor de Chi-cuadrado también es significativo frente al paso anterior ($p < 0{,}001$), y el modelo sigue siendo significativo ($p < 0{,}001$).

Paso 10. En el «Resumen de modelo» se nos indica el valor del estadístico «-2 log de la verosimilitud» (útil para hacer comparaciones) así como 2 valores de R cuadrado que no son equivalentes a r^2 de la regresión lineal, por lo que deben ser utilizados con cautela. El r^2 de Cox y Snell no varía entre 0 y 1 por lo que se ha modificado en lo que se llama el r^2 de Nagelkerke que sí varía entre 0 y 1.

CUADRO XI.4. *(Continuación)*[SPSS]

Atendiendo a nuestro caso, tendríamos que en el paso 3 se explica un 63,5% de la variabilidad de los datos.

Resumen del modelo

Paso		dimension1		
		-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
dimension0	1	779,735[a]	,321	,445
	2	619,249[b]	,434	,603
	3	581,820[b]	,457	,635

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

b. La estimación ha finalizado en el número de iteración 6 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Paso 11. La «Tabla de clasificación» nos indica el porcentaje de clasificaciones correctas con nuestro modelo. Se observa que las diferencias entre el paso 2 y el 3 son pequeñas. En el paso 3 el modelo predice correctamente el 74,7% de las personas que no tienen cáncer. Sin embargo, lo importante es que es capaz de predecir correctamente el 95,4% de las personas que tienen cáncer. Por lo tanto, es un modelo relativamente bueno para poder predecir cuando una persona potencialmente puede tener ese tipo cáncer.

Tabla de clasificación[a]

Observado					Pronosticado dimension2		
					Cáncer		Porcentaje correcto
					No	Si	
dimension0	Paso 1	dimension1	Cáncer	No	220	73	75,1
				Si	77	510	86,9
			Porcentaje global				83,0
	Paso 2	dimension1	Cáncer	No	212	81	72,4
				Si	24	563	95,9
			Porcentaje global				88,1
	Paso 3	dimension1	Cáncer	No	219	74	74,7
				Si	27	560	95,4
			Porcentaje global				88,5

a. El valor de corte es ,500

CUADRO XI.4. *(Continuación)*[SPSS]

Paso 12. La tabla «Variables en la ecuación» nos indica las variables seleccionadas en cada paso. En todos los pasos, todas las variables son significativas ($p < 0{,}001$).

El paso 1 incluye la variable más importante que es la presencia de sangre en la orina. Como el coeficiente (2,994) es positivo y ese coeficiente hace referencia a la segunda categoría (presencia de sangre en la orina), esto indica que las personas con cáncer suelen tener sangre en la orina. En el paso 2 se incorpora la edad al modelo. De nuevo el coeficiente positivo (0,067) indica que a mayor edad es más frecuente la aparición del cáncer. Por último, en el paso 3 se incorpora el estado civil de la persona. En el caso de que las variables categóricas no sean dicotómicas (sólo el estado civil), se crea un coeficiente B para cada categoría distinta de la de referencia (que tiene valor 0). El valor positivo del coeficiente B nos indica que entre las personas casadas (0,966) y las separadas (2,123) la presencia de cáncer es más frecuente. El hecho de que sea hombre o mujer no influye en la presencia del cáncer, ya que esta variable no se incorpora en ninguno de los pasos.

Especial importancia tiene en esta tabla el valor «Exp(B)», que es la OR (Odds Ratio, razón impar o razón de ventajas) que representa el cociente entre la probabilidad de que ocurra el suceso que define la variable dependiente frente a la probabilidad de que no ocurra en presencia o ausencia del factor. Así, el valor 38,646, correspondiente a la variable «Sangre.en.orina», nos indica que es 38,6 veces más grande el cociente entre la probabilidad de tener cáncer que de no tenerlo, en aquellas personas con sangre en la orina.

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Sangre.en.orina	2,994	,182	270,026	1	,000	19,961
	Constante	-1,050	,132	62,862	1	,000	,350
Paso 2[b]	Edad	,067	,006	110,755	1	,000	1,069
	Sangre.en.orina	3,502	,239	215,450	1	,000	33,184
	Constante	-4,436	,382	134,588	1	,000	,012
Paso 3[c]	Edad	,067	,007	102,915	1	,000	1,070
	Estado.Civil			34,228	2	,000	
	Estado.Civil(1)	,966	,229	17,748	1	,000	2,627
	Estado.Civil(2)	2,123	,396	28,717	1	,000	8,358
	Sangre.en.orina	3,654	,258	200,870	1	,000	38,646
	Constante	-5,318	,447	141,339	1	,000	,005

a. Variable(s) introducida(s) en el paso 1: Sangre.en.orina.

b. Variable(s) introducida(s) en el paso 2: Edad.

c. Variable(s) introducida(s) en el paso 3: Estado.Civil.

CUADRO XI.4. *(Continuación)*[SPSS]

Paso 13. El modelo vendría definido por la siguiente ecuación:

$$\Pi = \frac{1}{1 + e^{-(-5{,}318+0{,}067*Edad+0{,}966*(Casado)+2{,}123*(Separado)+3{,}654*(Con.sangre)}}$$

Por ejemplo, supongamos el primer caso de la tabla de datos, la probabilidad de que una persona casada, de 59 años, y con presencia de sangre en la orina tenga este tipo de cáncer sería:

$$\Pi = \frac{1}{1 + e^{-(-5{,}318+0{,}067*59+0{,}966*1+2{,}123*0+3{,}654*1}} = 0{,}96$$

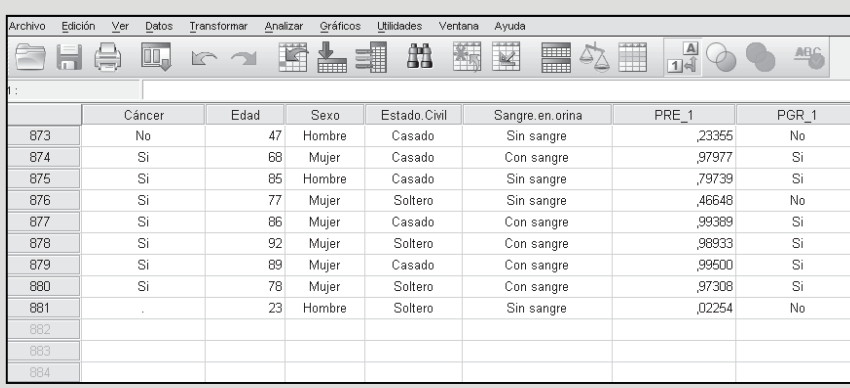
Como es mayor que 0,5, se pronostica que la persona tendría cáncer.

La probabilidad de que una persona soltera, de 23 años, y sin sangre en la orina, tenga este tipo de cáncer sería:

$$\Pi = \frac{1}{1 + e^{-(-5{,}318+0{,}067*23+0{,}966*0+2{,}123*0+3{,}654*0}} = 0{,}022$$

Como es menor que 0,5, se pronostica que la persona no tendría cáncer.

Paso 14. SPSS permite calcular las probabilidades y grupo de pertenencia. Para ello, simplemente se ponen al final de la tabla de datos, los casos que se quieren pronosticar, y en la ventana del paso 7 se marca «Probabilidades» y «Grupo de pertenencia»


Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana Ayuda

	Cáncer	Edad	Sexo	Estado.Civil	Sangre.en.orina	PRE_1	PGR_1
873	No	47	Hombre	Casado	Sin sangre	,23355	No
874	Si	68	Mujer	Casado	Con sangre	,97977	Si
875	Si	85	Hombre	Casado	Sin sangre	,79739	Si
876	Si	77	Mujer	Soltero	Sin sangre	,46648	No
877	Si	86	Mujer	Casado	Con sangre	,99389	Si
878	Si	92	Mujer	Soltero	Con sangre	,98933	Si
879	Si	89	Mujer	Casado	Con sangre	,99500	Si
880	Si	78	Mujer	Soltero	Con sangre	,97308	Si
881	.	23	Hombre	Soltero	Sin sangre	,02254	No
882							
883							
884							

Se ha puesto el ejemplo mostrado anteriormente, de una persona soltera de 23 años y sin sangre en la orina. Se obtiene una probabilidad de 0,022 y, por tanto, el pronóstico es que no tiene cáncer.

XI.2.2. REGRESIÓN LOGÍSTICA MULTINOMIAL

La diferencia principal con respecto a la regresión logística binomial es que la variable dependiente cualitativa no es dicotómica, sino que puede tener más de 2 categorías.

Suponiendo *k* clases o categorías, y *j* variables independientes, este modelo se puede resumir en las siguientes funciones, que proporcionan las probabilidades de pertenencia a las primeras *k* - 1 clases:

$$\Pi_{in} = \frac{e^{Z_{in}}}{1 + e^{Z_{i1}} + e^{Z_{i2}} + e^{Z_{i3}} + \ldots + e^{Z_{ik-1}}}$$

$$Z_{in} = \beta_{n0} + \beta_{n1}x_{i1} + \beta_{n2}x_{i2} + \ldots + \beta_{nj}x_{ij}$$

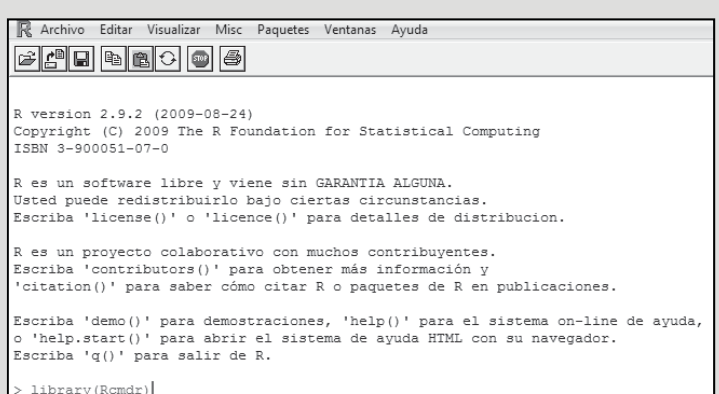
donde Π_{in} es la probabilidad de pertenencia del caso *i* a la clase *n*; Z_{in} es el valor de la variable dependiente *Z* correspondiente a la clase *n* en el caso *i*; β_{nj} es el coeficiente de la variable independiente *j* para la clase *n*; x_{ij} es el valor del predictor o variable independiente *j* para el caso *i*. La probabilidad para la última clase *k* se obtiene por diferencia a 1.

CUADRO XI.5. Regresión logística multinomial

Aplicación con R.

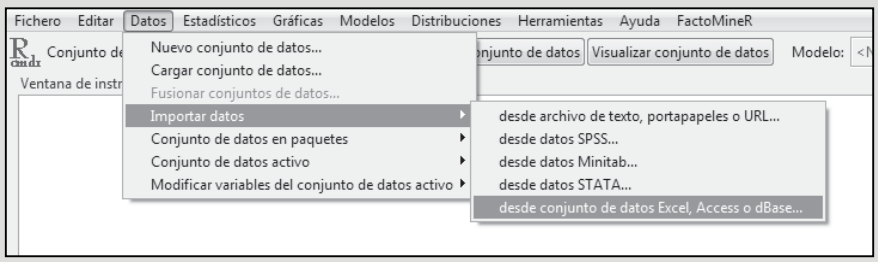
EJEMPLO. En el archivo de **Cuadro XI.5.xls** hay datos de medidas corporales de varias especies de peces (variables continuas), así como el tipo de boca que tienen (variable categórica). El objetivo es determinar si es posible identificar a que especie pertenece un individuo en función de la biometría. Sólo se va a trabajar con las familias Characidae, Curimatidae y Anostomidae.

Paso 1. Son necesarios los paquetes "nnet" y "Rcmdr". Si los paquetes no están instalados, la instalación se realiza como se explica en el Apéndice. Iniciamos R y cargamos el paquete «Rcmdr» introduciendo la orden «library(Rcmdr)».

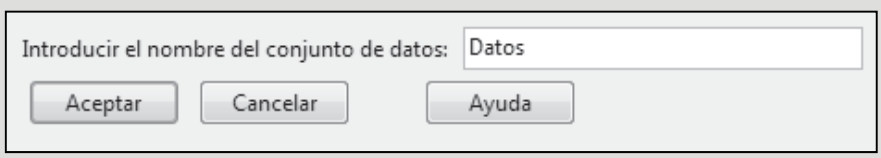

Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> library(Rcmdr)

CUADRO XI.5. *(Continuación)*[R]

Paso 2. Después de haber seleccionado el directorio de trabajo, como se explica en el Apéndice II, vamos a «Datos» ⇨ «Importar datos» ⇨ «desde conjunto de datos Excel, Access o dBase».


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda FactoMineR
Nuevo conjunto de datos...
Cargar conjunto de datos...
Fusionar conjuntos de datos...
Importar datos
Conjunto de datos en paquetes
Conjunto de datos activo
Modificar variables del conjunto de datos activo
desde archivo de texto, portapapeles o URL...
desde datos SPSS...
desde datos Minitab...
desde datos STATA...
desde conjunto de datos Excel, Access o dBase...
Visualizar conjunto de datos
Modelo:
Ventana de instr

Paso 3. Aparece un pequeño cuadro de diálogo para introducir el nombre del conjunto de datos, en el que dejamos el nombre por defecto «Datos» y pulsamos «Aceptar». Posteriormente aparecen los archivos del directorio de trabajo, en el que elegimos **Cuadro XI.5.xls** y pulsamos «Abrir».



Paso 4. En R Commander pulsamos en «Visualizar conjunto de datos» lo que nos permite inspeccionar nuestros datos.

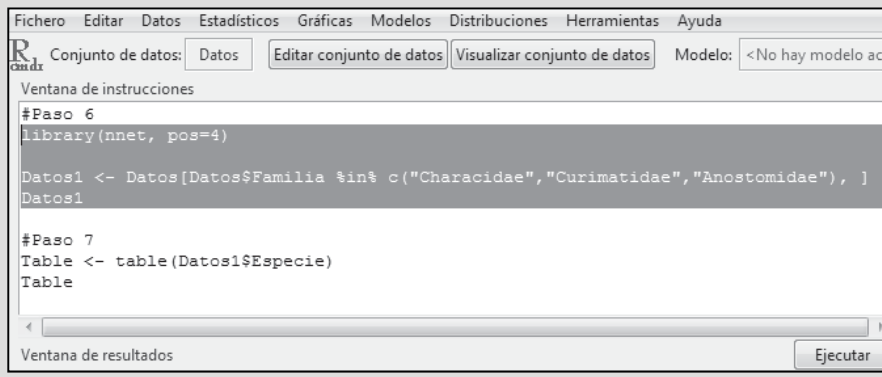
	Orden	Familia	Especie	M2	M6	M8	M9
1	Characiformes	Acestrorhynchidae	Acestrorhynchus abbreviatus	0.11	0.62	0.30	0.53
2	Characiformes	Acestrorhynchidae	Acestrorhynchus abbreviatus	0.09	0.59	0.27	0.49
3	Characiformes	Acestrorhynchidae	Acestrorhynchus abbreviatus	0.09	0.60	0.29	0.51
4	Characiformes	Acestrorhynchidae	Acestrorhynchus abbreviatus	0.09	0.60	0.28	0.51
5	Characiformes	Acestrorhynchidae	Acestrorhynchus abbreviatus	0.10	0.61	0.29	0.49
6	Characiformes	Acestrorhynchidae	Acestrorhynchus abbreviatus	0.10	0.60	0.29	0.50
7	Characiformes	Acestrorhynchidae	Acestrorhynchus abbreviatus	0.10	0.61	0.28	0.50
8	Characiformes	Acestrorhynchidae	Acestrorhynchus abbreviatus	0.09	0.60	0.27	0.50
9	Characiformes	Acestrorhynchidae	Acestrorhynchus abbreviatus	0.09	0.60	0.29	0.50
10	Characiformes	Acestrorhynchidae	Acestrorhynchus abbreviatus	0.12	0.60	0.32	0.54
11	Characiformes	Acestrorhynchidae	Acestrorhynchus abbreviatus	0.10	0.59	0.30	0.49
12	Characiformes	Acestrorhynchidae	Acestrorhynchus abbreviatus	0.11	0.60	0.32	0.52
13	Characiformes	Acestrorhynchidae	Acestrorhynchus abbreviatus	0.10	0.60	0.30	0.53
14	Characiformes	Acestrorhynchidae	Acestrorhynchus microlepis	0.11	0.66	0.27	0.52
15	Characiformes	Acestrorhynchidae	Acestrorhynchus microlepis	0.12	0.66	0.29	0.54
16	Characiformes	Acestrorhynchidae	Acestrorhynchus microlepis	0.12	0.66	0.29	0.51
17	Characiformes	Acestrorhynchidae	Acestrorhynchus microlepis	0.11	0.61	0.28	0.51

Paso 5. A continuación se carga el archivo de instrucciones **Cuadro XI.5.R** seleccionando «Fichero» ⇨ «Abrir fichero de instrucciones…» y no se guarda la salida.

Paso 6. La primera instrucción carga el paquete necesario y las siguientes seleccionan, del archivo de datos, las tres familias con las que se va a trabajar.

CUADRO XI.5. *(Continuación)*[R]

```
library(nnet, pos=4)
Datos1 <- Datos[Datos$Familia %in% c("Characidae","Curimatidae",
"Anostomidae"), ]
Datos1
```



Paso 7. Por medio de las siguientes instrucciones se obtiene la tabla de frecuencias de los datos. Se observa que hay 11 especies diferentes, con frecuencias que oscilan entre 7 y 14 (no se muestran todas en la ventana siguiente). El modelo debe expresar la probabilidad de que un elemento pertenezca a cada una de las clases como función de las variables explicativas, de modo que podremos identificar razonablemente la especie a partir de los valores de las variables independientes o explicativas.

```
Datos2 <- factor(Datos1$Especie, exclude=NULL) # elimina clases vacías
Table <- table(Datos2) # frecuencias de cada una de las clases
Table
```

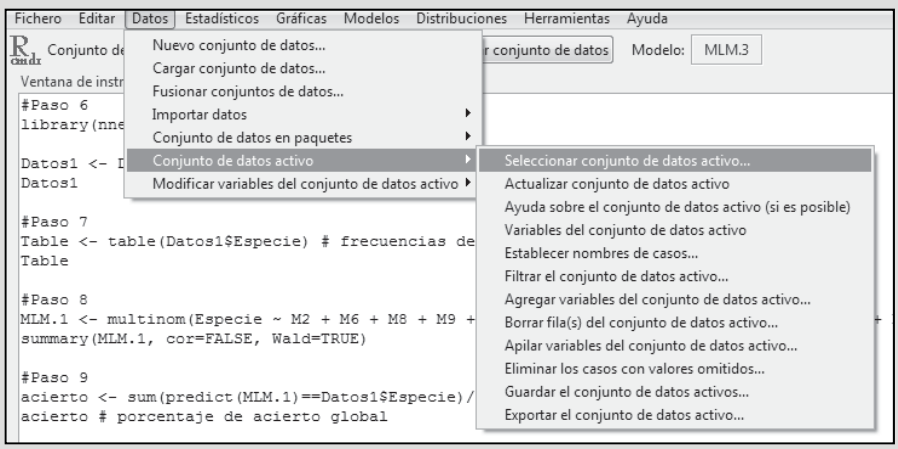
```
Datos2
      Bryconops melanurus           Charax michaeli        Curimata cisandina
                        7                         9                       14
       Roeboides myersii       Schizodon fasciatus Tetragonopterus argenteus
                        7                         7                        7
```

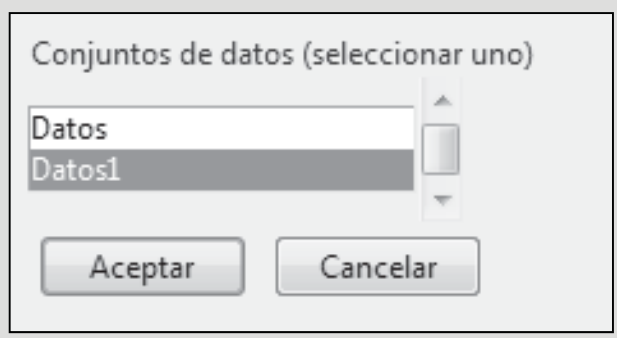
Paso 8. A continuación realizamos la regresión logística multinomial con la «Especie» como variable dependiente y considerando todas las variables independientes. El programa tarda varios minutos en construir el modelo. En el caso de que existan muchas clases en la variable dependiente y/o muchas variables independientes, es posible que el programa muestre un mensaje de error en el número máximo de "weights". No hay un límite intrínseco, pero aumentar el valor con la instrucción «MaxNWts=», poniendo un número mayor, probablemente permitirá ajustes que son muy lentos y que requieren mucho tiempo de ejecución.

```
MLM.1 <- multinom(Especie ~ M2 + M6 + M8 + M9 + M12 + M13 + M15
+ M16 + M17 + M21 + M22 + Boca, data=Datos1, trace=FALSE)
summary(MLM.1, cor=FALSE, Wald=TRUE)
```

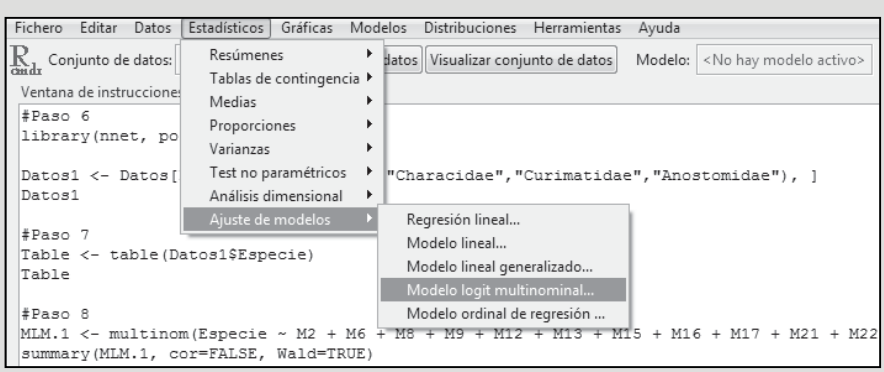
CUADRO XI.5. *(Continuación)*[R]

En vez de tener que escribir todas las variables, es posible seleccionarlas a través del menú. Para ello, en primer lugar vamos a «Conjunto» ⇨ «Seleccionar conjunto de datos activo…», para dejar como matriz para los análisis «Datos1» -la cual tiene la elección que acabamos de realizar de trabajar sólo con tres familias- en vez de la matriz «Datos».


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda
Nuevo conjunto de datos...
Cargar conjunto de datos...
Fusionar conjuntos de datos...
Importar datos
Conjunto de datos en paquetes
Conjunto de datos activo
Modificar variables del conjunto de datos activo
Seleccionar conjunto de datos activo...
Actualizar conjunto de datos activo
Ayuda sobre el conjunto de datos activo (si es posible)
Variables del conjunto de datos activo
Establecer nombres de casos...
Filtrar el conjunto de datos activo...
Agregar variables del conjunto de datos activo...
Borrar fila(s) del conjunto de datos activo...
Apilar variables del conjunto de datos activo...
Eliminar los casos con valores omitidos...
Guardar el conjunto de datos activos...
Exportar el conjunto de datos activo...
Modelo: MLM.3
#Paso 6
#Paso 7
Table <- table(Datos1$Especie) # frecuencias de
#Paso 8
MLM.1 <- multinom(Especie ~ M2 + M6 + M8 + M9 +
summary(MLM.1, cor=FALSE, Wald=TRUE)
#Paso 9
acierto <- sum(predict(MLM.1)==Datos1$Especie)/
acierto # porcentaje de acierto global

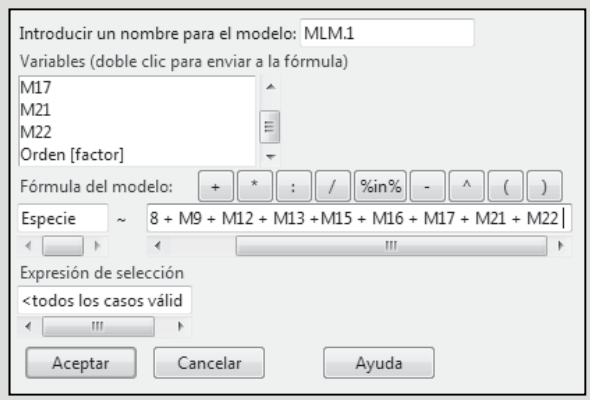


A continuación se selecciona «Estadísticos» ⇨ «Ajuste de modelos» ⇨ «Modelo logit multinomial».



CUADRO XI.5. *(Continuación)*[R]

En la ventana que aparece, se posiciona en el cuadro de las variables dependiente e independientes y se pulsa dos veces con el ratón sobre cada una de las variables.


Introducir un nombre para el modelo: MLM.1
Variables (doble clic para enviar a la fórmula)
M17
M21
M22
Orden [factor]
Fórmula del modelo:
+ * : / %in% - ^ ( )
Especie ~ 8 + M9 + M12 + M13 +M15 + M16 + M17 + M21 + M22
Expresión de selección
<todos los casos válid
Aceptar
Cancelar
Ayuda

Paso 9. Las siguientes instrucciones permiten calcular el grado de acierto del modelo de regresión, el cual es del 100%.

```
acierto <- sum(predict(MLM.1)==Datos2)/ dim(Datos1)[1]*100
acierto # porcentaje de acierto global
```

```
> acierto # porcentaje de acierto global
[1] 100
```

Paso 10. Para ver si alguna de las variables independientes no son necesarias en el modelo, se realiza una regresión por pasos.

```
MLM.2 <-stepwise(MLM.1, direction='backward/forward', criterion='AIC')
summary(MLM.2, cor=FALSE, Wald=TRUE)
```

Las ventanas siguientes muestran los pasos penúltimo (variables M9, M13, M15) y último (variables M13 y M15) del modelo de regresión por pasos.

```
Step:  AIC=80
Especie ~ M9 + M13 + M15

       Df     AIC
- M9   10  72.008
<none>     80.000
+ M17  10 100.000
+ M6   10 100.000
+ M8   10 100.000
+ M12  10 100.000
+ M21  10 100.000
+ M16  10 100.000
+ M22  10 100.000
+ Boca 10 100.000
+ M2   10 100.000
- M15  10 134.688
- M13  10 139.669
```

```
Step:  AIC=72.01
Especie ~ M13 + M15

       Df     AIC
<none>     72.008
+ M8   10  80.000
+ M9   10  80.000
+ M22  10  80.000
+ M17  10  80.000
+ M6   10  80.000
+ M21  10  86.864
+ M16  10  87.651
+ Boca 10  87.873
+ M2   10  90.342
+ M12  10  92.201
- M13  10 181.272
- M15  10 229.090
```

CUADRO XI.5. *(Continuación)*[R]

Paso 11. Se ejecutan las instrucciones de este paso, las cuales ya se explicaron en el paso 9, y se observa que el porcentaje global de acierto de este modelo 2 es un 97%.

```
acierto <- sum(predict(MLM.2)==Datos2)/ dim(Datos1)[1]*100
acierto # porcentaje de acierto global
```

```
> acierto # porcentaje de acierto global
[1] 97
```

Paso 12. Para intentar conseguir que se identifique de forma correcta el 100% de los casos, realizamos un modelo 3 incluyendo la variable M9 que se quitó en el último paso.

```
MLM.3 <- multinom(Especie ~ M9 + M13 + M15, data=Datos1, trace=FALSE)
summary(MLM.3, cor=FALSE, Wald=TRUE)
```

Paso 13. Se ejecutan las instrucciones de este paso que ya se explicaron en el paso 9, y se observa que ahora el porcentaje global de acierto de este modelo 3 es un 100%.

```
acierto <- sum(predict(MLM.3)==Datos2)/ dim(Datos1)[1]*100
acierto # porcentaje de acierto global
```

```
> acierto # porcentaje de acierto global
[1] 100
```

Paso 14. Las siguientes instrucciones generan el archivo **Salida Cuadro XI.5-1.csv**, en el cual hay las predicciones y las especies reales. Este archivo es útil para detectar fallos, cuando el porcentaje de acierto no sea del 100%.

	Predicción	Real
1	Rhytiodus microlepis	Rhytiodus microlepis
2	Rhytiodus microlepis	Rhytiodus microlepis
3	Rhytiodus microlepis	Rhytiodus microlepis

```
pred <- data.frame(predict(MLM.3),Datos1$Especie) # Se crea una matriz con dos columnas: predicciones y datos reales
names(pred)<-c("Predicción","Real") # se le cambia el nombre a las columnas para que se llamen «Predicción» y «Real»
write.csv2(pred,file="Salida Cuadro XI.5-1.csv") # Genera el archivo con las especies que predice el modelo y las reales
```

Paso 15. La siguiente instrucción permite guardar los coeficientes del modelo de regresión en el archivo **Salida Cuadro XI.5-2.csv**.

```
write.csv2(coefficients(MLM.3),file="Salida Cuadro XI.5-2.csv")
```

CUADRO XI.5. *(Continuación)*[R]

Una vez que se abre el archivo, se observan los coeficientes de las funciones estimadas para cada clase (en este caso la especie) excepto la primera (*Bryconops melanurus*), cuya probabilidad se obtiene por diferencia a 1.

	(Intercept)	M9	M13	M15
Charax michaeli	5248,71191	-379675,171	-30371,4238	424883,967
Curimata cisandina	-248090,137	157426,21	218524,814	197751,763
Curimatella meyeri	-150438,946	163473,204	218685,875	-100561,895

A modo de ejemplo se estima la probabilidad para el siguiente individuo:

Especie	M2	M6	M8	M9	M12	M13	M15	M16	M17	M21
Charax michaeli	0,07	0,51	0,27	0,37	0,13	0,42	0,46	0,38	0,34	0,28

Para estimar a qué especie pertenece un individuo, con los coeficientes del archivo **Salida Cuadro XI.5-2.csv** se calculan los valores *Z* para cada especie. Sólo se muestra como sería el cálculo de *Z* para dos especies: *Charax michaeli* y *Curimata cisandina*, realizando de la misma forma el cálculo de *Z* para el resto de especies.

Charax michaeli

$Z_{Ch.michaeli}$ = 5248,71 - 379675,17**M9* - 30371,4**M13* + 424883,97**M15*

$Z_{Ch.michaeli}$ = 5248,71 - 379675,17**0,37* - 30371,4**0,42* + 424883,97**0,46* = 47459

Curimata cisandina

$Z_{C.\ cisandina}$ = -248090,1 + 157426,2**M9* + 218524,8**M13* + 197751,7**M15*

$Z_{C.\ cisandina}$ = -248090,1 + 157426,2**0,37* + 218524,8**0,42* + 197751,7**0,46* = -7096,2

Una vez que se obtienen todos los valores *Z* para cada una de las especies, se calcula la probabilidad de pertenencia a cada especie. Es necesario mencionar que, al calcular el exponencial de *Z* con Excel, a veces es posible obtener un error si los valores son muy grandes, como ocurre en este ejemplo. En caso de obtener error con uno o más valores de *Z*, la mejor solución es restar a todos los valores *Z* una constante para que ninguno sea superior a 700. En este caso a todos los valores de *Z* se les quita 47000. A modo de ejemplo, sólo se muestra la probabilidad para la especie *Charax michaeli*, y se concluye que el individuo pertenece a esta especie con una probabilidad de 1.

$$\Pi_{Ch.michaeli} = \frac{e^{459,5}}{1 + e^{459,5} + e^{-54096,2} + e^{-91364,3} \ldots + e^{-92714,5}} = 1$$

CUADRO XI.5. *(Continuación)*[R]

Paso 16. En vez de tener que realizar los cálculos anteriores, es posible que R realice las predicciones. Para ello, tenemos en el archivo **Cuadro XI.5-1.V.csv** los datos de los individuos que se quiere conocer a qué especie pertenecen. Aunque no se especifica, el primer individuo es *Triportheus angulatus* y el segundo *Potamorhina altamazonica*, pero no importa si en el archivo se da la información de a que especie pertenece el individuo, porque R no utiliza esa información a la hora de realizar la predicción con el modelo de regresión.

Orden	Familia	Especie	M2	M6	M8	M9	M12	M13	M15	M16
Characiformes	Characidae		0,06	0,57	0,25	0,52	0,1	0,31	0,34	0,24
Characiformes	Curimatidae		0,07	0,45	0,32	0,51	0,15	0,42	0,31	0,3

```
DatosNuevos<-read.csv2("Cuadro XI.5-1.V.csv",enc="latin1") # abre el archivo con los datos
PredNuevos <- data.frame(predict(MLM.3,DatosNuevos),DatosNuevos$Especie) # crea una matriz en la que una columna son las predicciones y la otra las especies reales que se especifican en el archivo
names(PredNuevos)<-c("Predicción","Real") # se cambian los nombres de esta matriz para que una columna se llame «Predicción» y la otra «Real»
write.csv2(PredNuevos,file="Salida Cuadro XI.5-3.csv") # se guardan los resultados en un archivo
```

Al abrir el archivo **Salida Cuadro XI.5-3.csv** se observa que el modelo predice correctamente a que especie pertenece cada uno de los individuos.

	Predicción	Real
1	Triportheus angulatus	NA
2	Potamorhina altamazonica	NA

Paso 17. Por último, también podemos construir la matriz de clasificación, que indicará cuantos elementos de cada clase son asignados a esa misma clase y a cualquier otra. Los resultados se graban en dos archivos. En el archivo **Salida Cuadro XI.5-4.csv** están los elementos asignados de cada especie al conjunto de especies y en el archivo **Salida Cuadro XI.5-5.csv** se muestran los resultados en porcentaje. Estos archivos son útiles para detectar asignaciones erróneas, cuando el porcentaje de acierto no sea del 100%.

```
t <- table(predict(MLM.3),Datos2) # matriz de clasificación; filas: predicción; columnas: clase original excluyendo los datos nulos
tporc <- prop.table(t,2)*100 # matriz de clasificación con porcentajes de columna
write.csv2(t,file="Salida Cuadro XI.5-4.csv")
write.csv2(tporc,file="Salida Cuadro XI.5-5.csv")
```

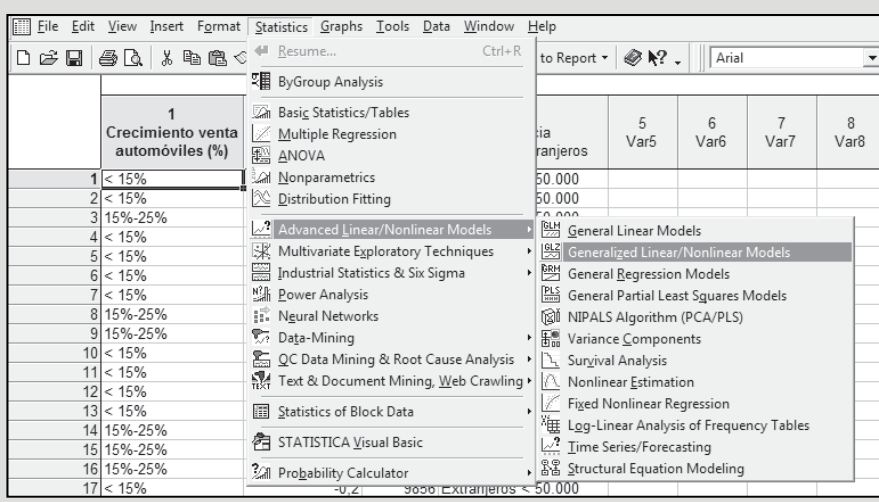
CUADRO XI.5. *(Continuación)*[STATISTICA]

Aplicación con STATISTICA.

EJEMPLO. En el archivo **Cuadro XI.5.sta** hay datos, para cada una de las provincias de España, del área en km^2 de cada provincia y, para el periodo 2004-2009, del crecimiento en la venta de automóviles (1 fue < 15%, 2 fue entre el 15%-25% y 3 fue > 25%), el porcentaje de crecimiento medio anual de la población y, por último, la diferencia entre lo que creció el número de españoles y extranjeros (1 creció más el número de españoles, 2 creció más el número de extranjeros en menos de 50.000 personas y 3 los extranjeros crecieron en más de 50.000 personas). Una empresa automovilística quiere determinar si es posible identificar las provincias con más crecimiento en ventas de automóviles a partir de las otras variables, para así poder obtener algún criterio que le permita tomar en el futuro la mejor decisión para ubicar nuevos concesionarios de venta.

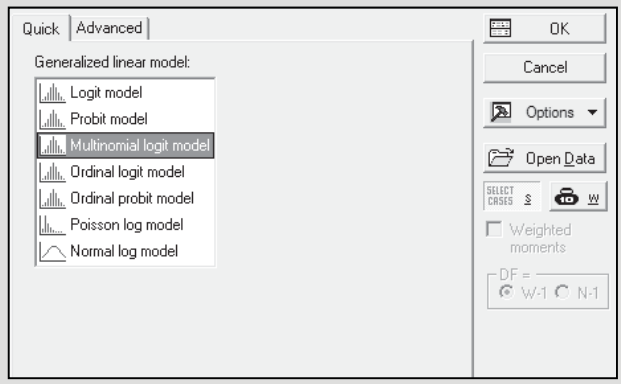
En ocasiones los datos complejos, en especial con muchas variables o muchos valores de la variable dependiente, presentan problemas matemáticos que afectan a los algoritmos utilizados por Statística para estimar los coeficientes del modelo de regresión, de modo que el programa emite un mensaje de error advirtiendo de la imposibilidad de llegar a un resultado. En estas situaciones puede ser aconsejable utilizar otro programa estadístico con menos limitaciones o simplificar el problema suprimiendo categorías de la variable dependiente y/o variables explicativas.

Paso 1. Se selecciona «Statistics» ⇨ «Advanced Linear/Nonlinear Models» ⇨ «Generalized Linear/Nonlinear Models».

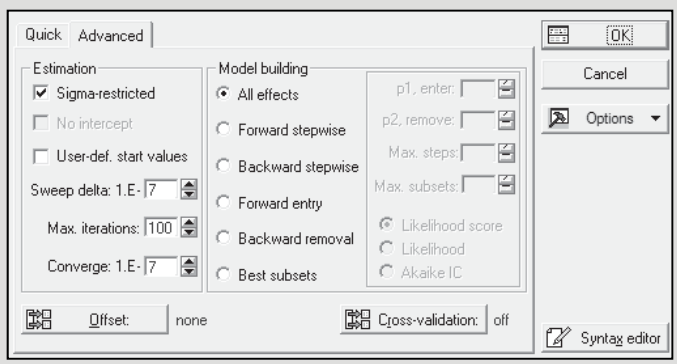


Paso 2. En la pestaña «Quick» se selecciona «Multinomial logit model».

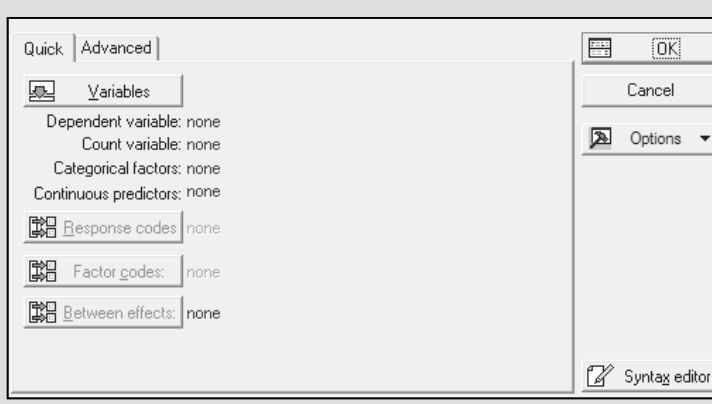
CUADRO XI.5. *(Continuación)* *STATISTICA*



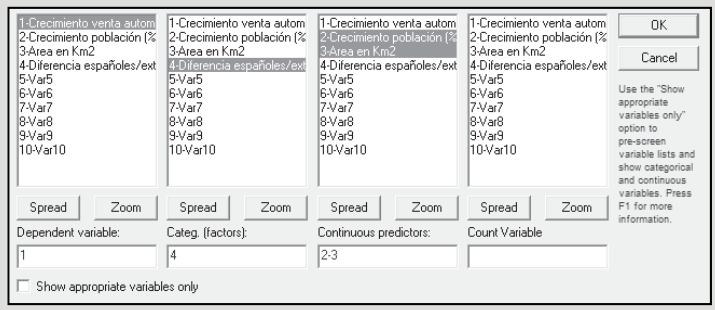
Paso 3. En la pestaña «Advanced» se elige «All effects», es decir, que se incluyan todas las variables independientes. Las otras formas de construir el modelo se explicaron en el paso 7 del Cuadro XI.2.



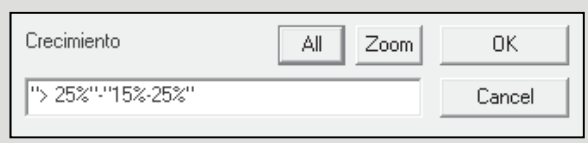
Paso 4. En la pestaña «Quick» se pulsa «Variables» y en la ventana que aparece se indica la variable dependiente, en este caso «Crecimiento venta automóviles (%)», y las independientes: «Diferencia españoles/extranjeros» que es una variable categórica y, «Area» y «Crecimiento población (%)» que son continuas.



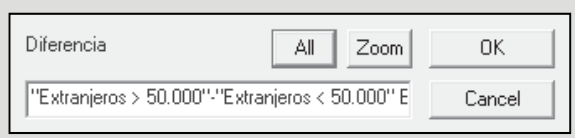
CUADRO XI.5. *(Continuación)* STATISTICA



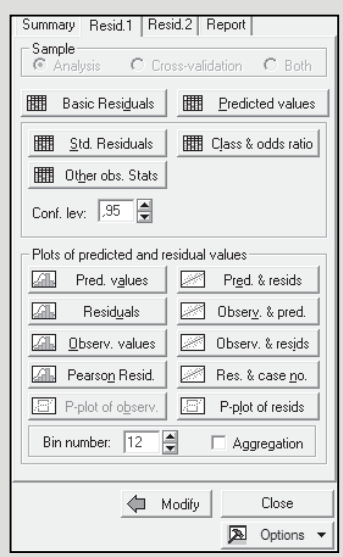
Paso 5. Para especificar las etiquetas de la variable dependiente, en la ventana del paso 4 se pulsa en «Response codes» y en la ventana que aparece se pulsa «All».



Para especificar las etiquetas de las variables independientes, en la ventana del paso 4 se pulsa en «Factor codes» y en la ventana que aparece se pulsa también «All».



Paso 6. Después de pulsar «OK» en la ventana del paso 4, en la siguiente ventana, en la pestaña «Resid.1», se pulsa «Class & odds ratio».

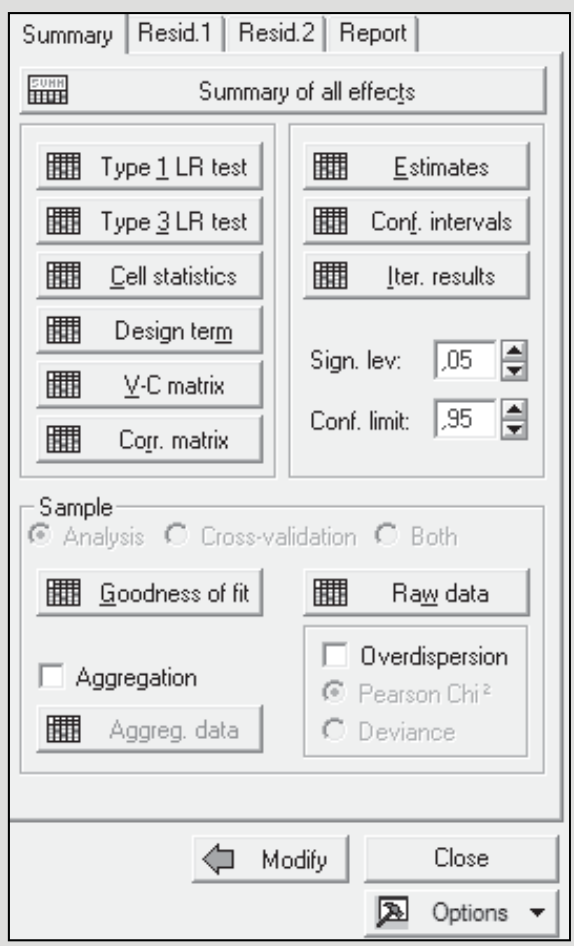


CUADRO XI.5. *(Continuación)*[STATISTICA]

Paso 7. Se observa que el modelo es capaz de predecir correctamente el 91,67% de las provincias donde el incremento en la venta de automóviles fue superior al 25%, pero solo el 66,67% de las provincias donde el crecimiento fue inferior al 15%.

	Classification of cases (Cuadro XI.5.sta) Odds ratio: 132,000000 Log odds ratio: 4,882802			
Observed	Predicted > 25%	Predicted < 15%	Predicted 15%-25%	Percent correct
> 25%	11	0	1	91,66667
< 15%	1	8	3	66,66667
15%-25%	2	2	24	85,71429

Paso 8. En la pestaña «Summary» se pulsa «Summary of all effects».



CUADRO XI.5. *(Continuación)* STATISTICA

Paso 9. Para cada variable en el modelo se muestra el estadístico de Wald, que es el cuadrado del cociente entre el valor estimado del coeficiente y su desviación típica, y que se utiliza para contrastar la hipótesis de que el coeficiente correspondiente es nulo, ya que en ese supuesto sigue una distribución Chi-cuadrado. Si el valor *p* es menor que el nivel de significación elegido (por ejemplo 0,05), debemos conservar esa variable en el modelo. En este ejemplo vemos que la variable «Diferencia españoles/extranjeros» no es significativa y se podría eliminar del modelo. Sin embargo, para explicar cómo se introducen variables categóricas en el modelo de regresión, no eliminaremos esta variable.

	Crecimiento venta automóviles (%) - Test of all effects (Cuadro XI.5.sta) Distribution : MULTINOMIAL Link function: LOGIT		
Effect	Degr. of Freedom	Wald Stat.	p
Intercept	2	5,074240	0,079094
Crecimiento población (%)	2	9,293746	0,009592
Area en Km2	2	7,444196	0,024183
Diferencia españoles/extranjeros	4	5,304230	0,257481

Paso 10. En la ventana del paso 8 se pulsa «Estimates». Se obtienen los coeficientes de la regresión. Hay una relación significativa positiva entre el área de la provincia y el grupo con incremento superior al 25% en la venta de automóviles. Por el contrario, el hecho de que el incremento de la población se deba en su mayoría a los extranjeros (>50.000 personas), influye negativamente sobre la probabilidad de que las ventas se incrementen por encima del 25%. El último coeficiente significativo hace referencia a que el crecimiento alto de la población reduce la probabilidad de que la subida de las ventas de automóviles sea inferior al 15%.

	Crecimiento venta automóviles (%) - Parameter estimates (Cuadro XI.5.sta) Distribution : MULTINOMIAL Link function: LOGIT						
Effect	Level of Effect	Level of Response	Column	Estimate	Standard Error	Wald Stat.	p
Intercept 1		> 25%	1	-8,81750	3,914	5,074231	0,024284
Crecimiento población (%)		> 25%	2	-0,51288	1,069	0,230387	0,631237
Area en Km2		> 25%	3	0,00077	0,000	6,874373	0,008744
Diferencia españoles/extranjeros	Extranjeros > 50.000	> 25%	4	-5,13748	2,257	5,181444	0,022829
Diferencia españoles/extranjeros	Extranjeros < 50.000	> 25%	5	-1,14967	1,232	0,870295	0,350874
Intercept 2		< 15%	6	-3,59043	1072,016	0,000011	0,997328
Crecimiento población (%)		< 15%	7	-2,19351	0,720	9,290287	0,002304
Area en Km2		< 15%	8	-0,00006	0,000	0,264376	0,607130
Diferencia españoles/extranjeros	Extranjeros > 50.000	< 15%	9	5,85785	1072,016	0,000030	0,995640
Diferencia españoles/extranjeros	Extranjeros < 50.000	< 15%	10	5,56831	1072,016	0,000027	0,995856
Scale				1,00000	0,000		

Por lo tanto, las ventas de automóviles son mayores en las provincias más grandes, y donde la población que más creció fue la de españoles; el incremento de las ventas es menor donde el crecimiento de la población fue pequeño.

CUADRO XI.5. *(Continuación)*[STATISTICA]

Paso 11. Para estimar si una provincia estaría en el grupo de crecimiento < 15%, entre 15%-25% ó > 25%, con los coeficientes mostrados en la tabla anterior «Estimate», se calculan los valores *Z* para cada grupo. El grupo de 15%-25% se estima por diferencia a 1 de la suma de las probabilidades obtenidas para los otros dos grupos.

Grupo > 25%

$Z_{>25\%}$ = -8,8175 - 0,5128**Población* + 0,00077**Area* - 5,1374**(> 50.000)* - 1,1496**(<50.000)*

Grupo < 15%

$Z_{<15\%}$ -3,5904 - 2,1935**Población* - 0,00006**Area* + 5,8578**(> 50.000)* + 5,5683**(<50.000)*

A modo de ejemplo se estima la probabilidad para el primer individuo de la matriz de datos.

	1 Crecimiento venta automóviles (%)	2 Crecimiento población (%)	3 Area en Km2	4 Diferencia españoles/extranjeros
1	< 15%	0,3	2217	Extranjeros > 50.000

$Z_{>25\%}$ = -8,8175 - 0,5128**0,3* + 0,00077**2217* - 5,1374**1* - 1,1496**0* = -12,39

$Z_{<15\%}$ -3,5904 - 2,1935**0,3* - 0,00006**2217* + 5,8578**1* + 5,5683**0* = 1,48

Por último, la probabilidad para cada grupo sería:

$$\Pi_{>25\%} = \frac{e^{-12,39}}{1 + e^{-12,39} + e^{1,48}} = 0 \qquad \Pi_{<15\%} = \frac{e^{1,48}}{1 + e^{-12,39} + e^{1,48}} = 0,81$$

Por lo tanto, el modelo predice que la primera provincia pertenece al grupo de crecimiento < 15% con una probabilidad de 0,81, que coincide con el grupo real de crecimiento de la provincia.

Paso 12. En la ventana del paso 6 se pulsa «Predicted values» y se obtienen las probabilidades que predice el modelo para cada uno de los casos y el intervalo de confianza al 95%.

	Crecimiento venta automóviles (%) - Predicted Values (Cuadro XI.5.sta) Distribution : MULTINOMIAL Link function: LOGIT											
Case	Response '1' = "> 25%"	Pred. Value	LINEAR Pred.	Standard Error	Lower CL 95, %	Upper CL 95, %	Response '1' = "< 15%"	Pred. Value	LINEAR Pred.	Standard Error	Lower CL 95, %	Upper CL 95, %
1	0,000000	0,000001	-12,3907	5,059430	0,000001	0,000001	1,000000	0,815093	1,4835	1,477	0,749341	0,880846
2	0,000000	0,000001	-12,7833	5,085798	0,000001	0,000001	1,000000	0,699074	0,8429	1,415	0,576307	0,821841
3	0,000000	0,000175	-8,2863	3,223434	0,000175	0,000175	0,000000	0,305157	-0,8226	0,865	0,228960	0,381353
4	0,000000	0,007452	-4,6961	1,923346	0,007246	0,007658	1,000000	0,176387	-1,5319	0,674	0,148502	0,204271
5	0,000000	0,003768	-3,6757	2,104383	0,003710	0,003826	1,000000	0,847473	1,7399	0,829	0,820317	0,874629

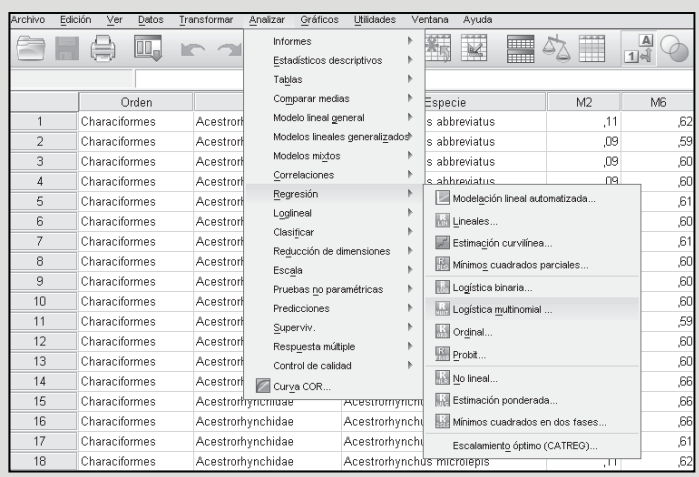
CUADRO XI.5. *(Continuación)*[SPSS]

Aplicación con SPSS.

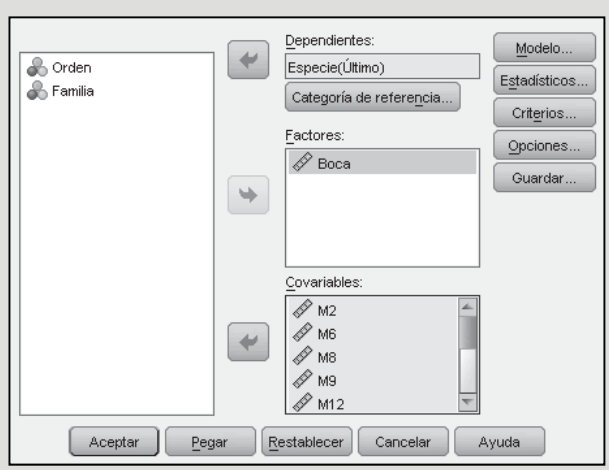
EJEMPLO. Se va utilizar el mismo ejemplo que en la aplicación con R. En el archivo **Cuadro XI.5.sav** hay datos de medidas corporales de varias especies de peces (variables continuas), así como el tipo de boca que tienen (variable categórica). El objetivo es determinar si es posible identificar a que especie pertenece un individuo en función de la biometría.

El proceso de etiquetado de variables se hace como se explicó en la regresión logística binaria (Cuadro XI.4, véase pasos 1 y 2).

Paso 1. Se selecciona «Analizar» ⇨ «Regresión» ⇨ «Logística multinomial…».

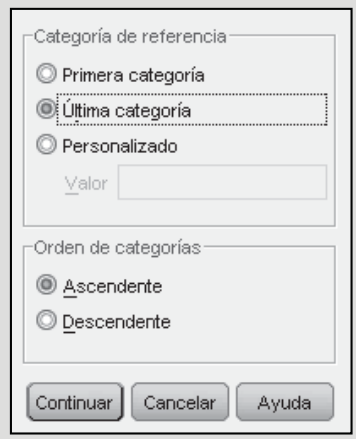


Paso 2. Se especifica cuál es la variable «Dependiente», en «Factores:» se incluyen las variables independientes categóricas y en «Covariables:» se ponen las variables independientes continuas.



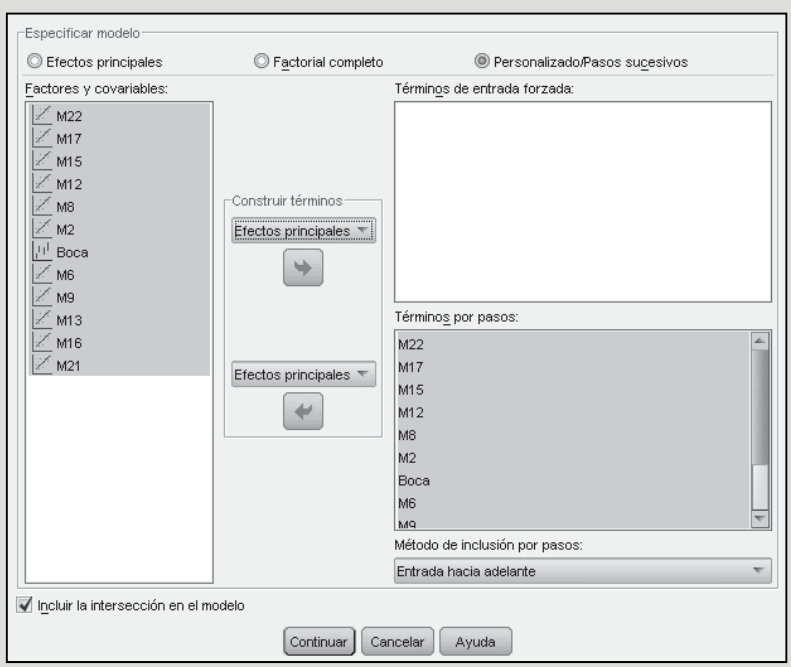
CUADRO XI.5. *(Continuación)*[SPSS]

Paso 3. En la variable dependiente aparece entre paréntesis la palabra «Último». Este paréntesis indica cual es la categoría de referencia. SPSS estimará los parámetros para las restantes. Pulsando sobre el icono «Categoría de referencia…» podemos elegir otra categoría de referencia.



Podemos elegir que la categoría de referencia sea la «Primera categoría», la «Última categoría» o «Personalizado» (indicar qué categoría queremos que sea). También es posible ordenar las categorías de forma «Ascendente» o «Descendente». Se escogen las opciones por defecto.

Paso 4. En primer lugar se hará un modelo por pasos. Para ello, en la ventana del paso 2, se pulsa «Modelo» y aparece la siguiente ventana.



CUADRO XI.5. *(Continuación)*[SPSS]

En el apartado «Especificar modelo» elegimos los términos que queremos incluir. Así tenemos 3 tipos de modelo:

-*Efectos principales.* No incluye las interacciones entre las distintas variables independientes.

-*Factorial completo.* Incluye todas las interacciones entre las distintas variables independientes, excepto entre las covariables.

-*Personalizado/Pasos sucesivos*. Permite especificar las interacciones que deseamos incluir, o solicitar un modelo por pasos (eligiendo sólo las variables significativas).

Se selecciona el modelo «Personalizado/Pasos sucesivos» y se activan las opciones para seleccionar el tipo de interacciones.

En «Construir términos» se puede seleccionar:

-*Interacción:* Incluye la interacción seleccionada.

-*Efectos principales:* Incluye los efectos principales de las variables.

-*Todas de 2:* Incluye todas las interacciones de dos variables.

-*Todas de 3:* incluye todas las interacciones de tres variables.

-*Todas de 4:* Incluye todas las interacciones de cuatro variables.

-*Todas de 5:* Incluye todas las interacciones de cinco variables.

Se selecciona «Efectos principales» tanto en «Términos de entrada forzosa:» como en «Términos por pasos:».

En la columna de la izquierda aparecen las variables independientes «Factores y Covariables», que deben pasarse a los cuadros de la derecha. Para cada variable independiente tenemos dos posibles opciones, ya que cada variable se puede introducir de forma forzosa en el modelo «Términos de entrada forzosa:» o introducirse únicamente si son estadísticamente significativas en el modelo «Términos por pasos:». En nuestro caso introduciremos todas las variables en el cuadro de «Términos por pasos:», como se muestra en la pantalla anterior, ya que queremos quedarnos únicamente con las variables significativas.

Una vez seleccionadas las variables debemos elegir el tipo de pasos sucesivos que queremos realizar en «Método de inclusión por pasos:». Existen 4 posibilidades:

-*Entrada hacia delante.* En cada paso se añade al modelo el término más significativo, hasta que ninguno de los términos por pasos que quede fuera del modelo tenga una contribución estadísticamente significativa si se añadiese al modelo.

CUADRO XI.5. *(Continuación)*[SPSS]

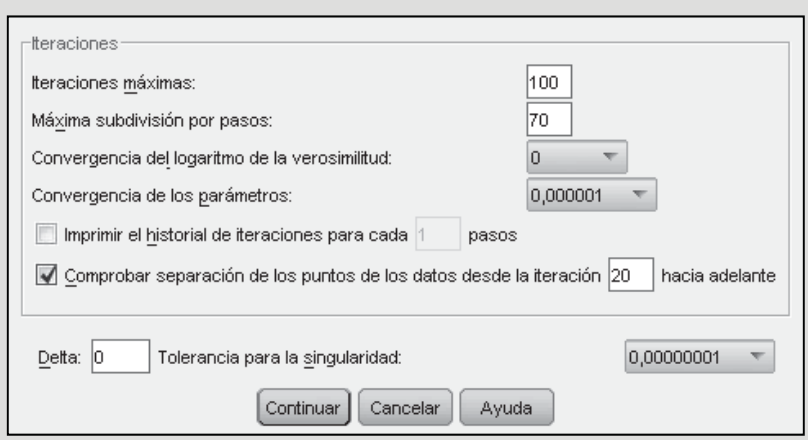
-Eliminación hacia atrás. Se inicia introduciendo en el modelo todos los términos especificados en la lista por pasos. En cada paso se elimina del modelo el término menos significativo, hasta que todos los términos por pasos restantes representen una contribución estadísticamente significativa para el modelo.

-Pasos sucesivos hacia adelante. Este método se inicia con el modelo que se seleccionaría mediante el método de entrada hacia delante. A partir de ahí, el algoritmo alterna entre la eliminación hacia atrás de los términos por pasos del modelo, y la entrada hacia delante de los términos fuera del modelo. Se sigue así hasta que no queden términos que cumplan con los criterios de entrada o exclusión.

-Pasos sucesivos hacia atrás. Este método se inicia con el modelo que se seleccionaría mediante el método de eliminación hacia atrás. A partir de ahí, el algoritmo alterna entre la entrada hacia delante de los términos fuera del modelo, y la eliminación hacia atrás de los términos por pasos del modelo. Se sigue así hasta que no queden términos que cumplan con los criterios de entrada o exclusión.

Se selecciona «Entrada hacia adelante».

Paso 5. En la pantalla principal (paso 2) pulsamos sobre el icono «Criterios».



En esta ventana podemos modificar las siguientes opciones (sólo se modifica la opción «Máxima subdivisión por pasos»):

- *Iteraciones máximas* de los algoritmos del modelo.

- *Máxima subdivisión por pasos*. Se cambia a 70.

- *La tolerancia de convergencia del logaritmo de la verosimilitud*. Se asume convergencia si el cambio absoluto es menor que el valor especificado (no negativo), en caso de ser 0 no se aplica este criterio.

CUADRO XI.5. *(Continuación)*[SPSS]

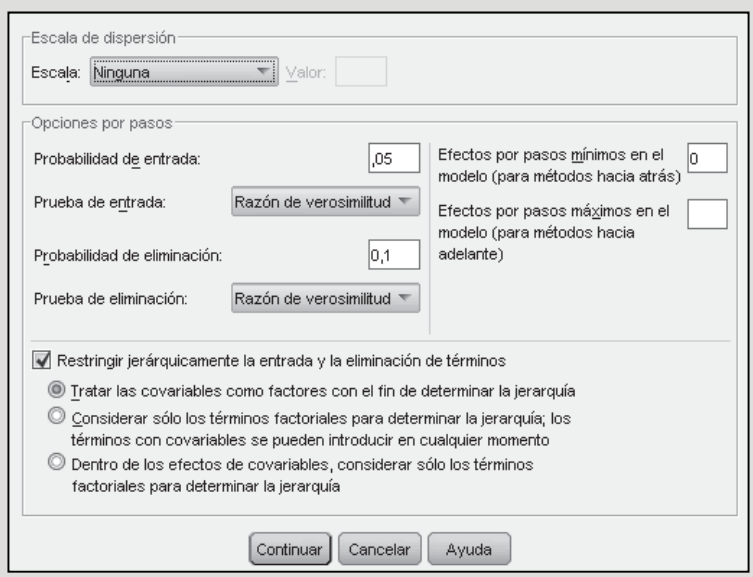
- *La tolerancia de convergencia de los parámetros*. Se asume convergencia si el cambio absoluto en las estimaciones de los parámetros es menor que el valor especificado (no negativo), en caso de ser 0 no se aplica este criterio.

- Imprimir el *historial de las iteraciones* para cada paso.

- Indicar *Delta* (entre 0 y 1). Se añade a cada casilla vacía de la tabla de contingencia, ayudando a estabilizar el algoritmo y evitar sesgos en las estimaciones.

- Indicar la *tolerancia para la singularidad*. Si algún elemento del modelo tiene una tolerancia menor, se excluye.

Paso 6. En la pantalla principal (paso 2) pulsamos sobre el icono «Opciones».


Escala de dispersión
Escala: Ninguna
Valor:
Opciones por pasos
Probabilidad de entrada: ,05
Prueba de entrada: Razón de verosimilitud
Probabilidad de eliminación: 0,1
Prueba de eliminación: Razón de verosimilitud
Efectos por pasos mínimos en el modelo (para métodos hacia atrás) 0
Efectos por pasos máximos en el modelo (para métodos hacia adelante)
Restringir jerárquicamente la entrada y la eliminación de términos
Tratar las covariables como factores con el fin de determinar la jerarquía
Considerar sólo los términos factoriales para determinar la jerarquía; los términos con covariables se pueden introducir en cualquier momento
Dentro de los efectos de covariables, considerar sólo los términos factoriales para determinar la jerarquía
Continuar
Cancelar
Ayuda

Las distintas opciones que podemos indicar son (no se modifican las opciones que el programa tiene por defecto):

-*Escala de dispersión*. Especifica el valor de escalamiento de la dispersión que se va a utilizar para corregir la estimación de la matriz de covarianzas de los parámetros. Existen tres tipos: *Desvianza* (mediante el estadístico de la función de desviación-chi-cuadrado de la razón de verosimilitud), *Pearson* (mediante el estadístico Chi-cuadrado de Pearson) y también se puede especificar su propio valor de escalamiento. Debe ser un valor numérico positivo.

- *Opciones por pasos*. Sólo si se utilizan métodos por pasos para generar un modelo. Las opciones que podemos modificar son:

CUADRO XI.5. *(Continuación)*[SPSS]

· *Probabilidad de entrada.* Con mayor probabilidad especificada, más fácil resultará que una variable entre en el modelo.

· *Prueba de entrada.* Método para introducir los términos en los métodos por pasos. Escoge entre la prueba de la razón de verosimilitud y la prueba de puntuación.

· *Probabilidad de eliminación.* En los métodos por pasos. Cuanto mayor sea la probabilidad especificada, más fácil resultará que una variable permanezca en el modelo.

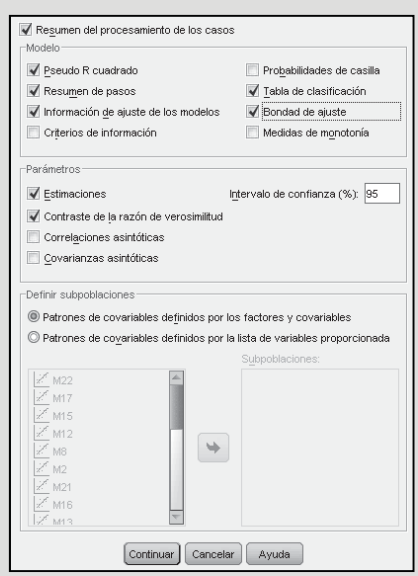
· *Prueba de eliminación.* Éste es el método utilizado para eliminar términos en los métodos por pasos. Puede elegir entre la prueba de la razón de verosimilitud o la prueba de Wald.

· *Efectos por pasos mínimos en el modelo* (para métodos hacia atrás). Mínimo número de términos que pueden incluirse en el modelo.

· *Efectos por pasos máximos en el modelo* (para métodos hacia adelante). Especifica el máximo número de términos que pueden incluirse en el modelo.

- *Restringir jerárquicamente la entrada y la eliminación de términos.* Permite aplicar restricciones a la inclusión de términos en el modelo. La jerarquía precisa que para que se incluya un término, todos los inferiores que formen parte del que se desea incluir, se encuentren antes en el modelo.

Paso 7. En la pantalla principal (paso 2) pulsamos sobre el icono «Estadísticos».



CUADRO XI.5. *(Continuación)*[SPSS]

Los estadísticos son:

-*Resumen de procesamiento de los casos.* Información sobre las variables categóricas especificadas.

-En la sección «Modelo»*:*

·*Pseudo R cuadrado.* Calcula el estadístico de Cox y Snell, de Nagelkerke y el r^2 McFadden. De estos tres *pseudo-*r^2 (imitación del r^2 de Pearson) únicamente el estadístico de Nagelkerke varía de 0 a 1.

·*Resumen de pasos.* Indica las variables introducidas o eliminadas.

·*Información de ajuste de los modelos.* Compara el modelo generado frente a la ausencia de modelo.

·*Criterios de información.* Indica el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC); basados en la "Teoría de la información" ayudan a seleccionar el mejor modelo, que es el aquel que tenga un menor valor en estos criterios.

·*Probabilidades de casilla.* Devuelve las frecuencias observadas y esperadas (con los residuos) y las proporciones por patrón en las covariables y por categoría de respuesta.

·*Tabla de clasificación.* Tabla de las respuestas observadas respecto a las respuestas pronosticadas. Un alto porcentaje de acierto indica que la regresión es adecuada.

·*Bondad de ajuste.* Indica los estadísticos de Chi-cuadrado de Pearson y de Chi-cuadrado de la razón de verosimilitud.

- En la sección «Parámetros»: Estadísticos relativos a los parámetros del modelo.

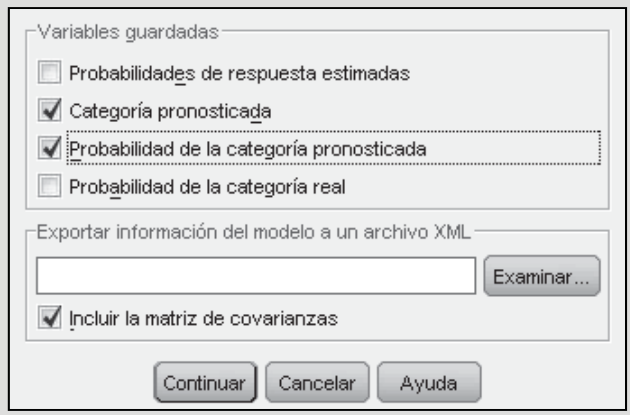
·*Estimaciones.* Parámetros del modelo con un nivel de confianza especificado.

·*Contraste de la razón de verosimilitud.* Razón de verosimilitud para los efectos parciales del modelo.

·*Correlaciones asintóticas.* Matriz de las correlaciones entre las estimaciones de los parámetros.

·*Covarianzas asintóticas*. Matriz de las covarianzas de las estimaciones de los parámetros.

CUADRO XI.5. *(Continuación)*[SPSS]



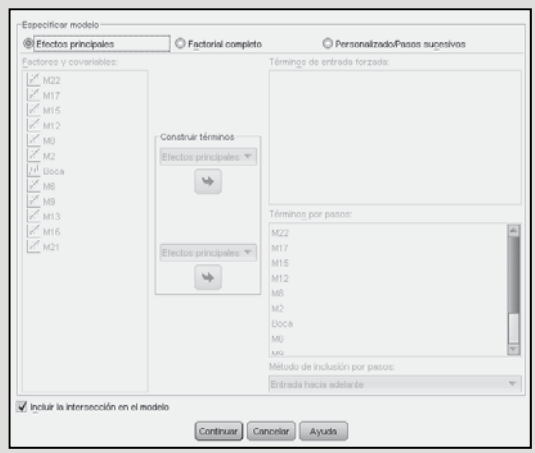
En nuestro caso seleccionamos «Categoría pronosticada» y «Probabilidad de la categoría pronosticada».

Paso 9. En primer lugar nos fijamos en la tabla de clasificación. Se observa que el modelo por pasos no ha sido capaz de escoger un buen modelo predictivo, ya que solo consigue predecir de forma correcta el 53,8% de los casos. Esto ocurre de forma frecuente cuando el número de clases de la variable dependiente es grande; en esa situación, a menudo las variables explicativas que tienen una mayor capacidad predictiva son distintas para cada clase, sin que exista un número reducido de variables comunes a todas las clases. En esos casos los datos presentan problemas matemáticos en la aplicación del algoritmo de estimación por el método de máxima verosimilitud que impiden a SPSS obtener una solución satisfactoria con la selección automática de variables. En lugar de utilizar un método paso a paso, incluiremos por lo tanto todas las variables en el modelo.

	Acestrorhynchus abbreviatus	Acestrorhynchus microlepis	Epapterus dispilurus	Hypoctopoma gulare	Pimelodus blochii	Porcentaje correcto
Acestrorhynchus abbreviatus	9	0	0	0	0	69,2%
Acestrorhynchus microlepis	2	0	0	0	0	,0%
Rhytiodus microlepis	0	0	0	0	0	50,0%
Schizodon fasciatus	0	0	0	0	0	,0%
Bryconops melanurus	0	0	0	1	0	,0%
Charax michaeli	0	0	0	2	0	,0%
Mylossoma aureum	0	0	0	0	0	28,6%
Mylossoma duriventre	0	0	0	0	0	84,6%
Roeboides myersii	0	0	0	2	0	,0%
Tetragonopterus argenteus	0	0	0	0	0	71,4%
Triportheus angulatus	6	0	0	0	0	,0%
Boulengerella maculata	0	0	0	0	0	100,0%
Curimata cisandina	0	0	0	0	0	78,6%
Curimatella meyeri	0	0	0	0	0	,0%
Potamorhina altamazonica	0	0	0	0	0	28,6%
Psectrogaster rutiloides	0	0	0	0	0	50,0%

CUADRO XI.5. *(Continuación)*[SPSS]

Paso 10. Para ello, en la ventana del paso 2, se pulsa «Modelo» y en la ventana que aparece se marca «Efectos principales». Con ello se elimina la opción por pasos y todas las variables independientes introducidas en el paso 2 se considerarán dentro del modelo.



Paso 11. De nuevo nos fijamos en la tabla de clasificación. Esta vez se observa que el modelo de regresión predice correctamente el 100% de los casos, por lo tanto, es posible averiguar a que especie pertenece un individuo con los datos biométricos.

	Acestrorhynchus abbreviatus	Acestrorhynchus microlepis	Rhytiodus microlepis	Epapterus dispilurus	Hypoctopoma gulare	Pimelodus blochii	Porcentaje correcto
Acestrorhynchus abbreviatus	13	0	0	0	0	0	100,0%
Acestrorhynchus microlepis	0	6	0	0	0	0	100,0%
Rhytiodus microlepis	0	0	14	0	0	0	100,0%
Schizodon fasciatus	0	0	0	0	0	0	100,0%
Bryconops melanurus	0	0	0	0	0	0	100,0%
Charax michaeli	0	0	0	0	0	0	100,0%
Mylossoma aureum	0	0	0	0	0	0	100,0%
Mylossoma duriventre	0	0	0	0	0	0	100,0%
Roeboides myersii	0	0	0	0	0	0	100,0%
Tetragonopterus argenteus	0	0	0	0	0	0	100,0%
Triportheus angulatus	0	0	0	0	0	0	100,0%
Boulengerella maculata	0	0	0	0	0	0	100,0%
Curimata cisandina	0	0	0	0	0	0	100,0%
Curimatella meyeri	0	0	0	0	0	0	100,0%
Potamorhina altamazonica	0	0	0	0	0	0	100,0%
Psectrogaster rutiloides	0	0	0	0	0	0	100,0%
Rhaphiodon vulpinus	0	0	0	0	0	0	100,0%
Hoplias malabaricus	0	0	0	0	0	0	100,0%
Carnegiella strigata	0	0	0	0	0	0	100,0%
Anodus elongatus	0	0	0	0	0	0	100,0%
Hemiodus microlepis	0	0	0	0	0	0	100,0%
Prochilodus nigricans	0	0	0	0	0	0	100,0%
Semaprochilodus insignis	0	0	0	0	0	0	100,0%
Cichla monoculus	0	0	0	0	0	0	100,0%
Satanoperca jurupari	0	0	0	0	0	0	100,0%
Plagioscion squamosissimus	0	0	0	0	0	0	100,0%
Epapterus dispilurus	0	0	0	7	0	0	100,0%
Hypoctopoma gulare	0	0	0	0	14	0	100,0%
Pimelodus blochii	0	0	0	0	0	7	100,0%
Porcentaje global	4,7%	2,2%	5,0%	2,5%	5,0%	2,5%	100,0%

CUADRO XI.5. *(Continuación)*[SPSS]

Paso 12. Es posible que algunas de las variables no sean necesarias para llegar a predecir correctamente el 100% de los casos, es decir, que sean redundantes o no ayuden a discriminar las diferentes clases de la variable dependiente. Para eliminar aquellas variables que no sean útiles en el modelo nos fijamos en la tabla «Contrastes de la razón de verosimilitud». Lo mejor es ir eliminando de una en una las variables que no son significativas ($p > 0,05$), comenzando por la que tenga un valor de Criterio de ajuste más bajo. En este caso debería ser la variable categórica posición de boca, pero para explicar cómo introducir variables categóricas en el modelo, la dejaremos y eliminamos la siguiente con valor más bajo que sería la M2 (2,695).

Contrastes de la razón de verosimilitud

Efecto	Criterio de ajuste del modelo	Contrastes de la razón de verosimilitud		
	-2 log verosimilitud del modelo reducido	Chi-cuadrado	gl	Sig.
Intersección	4,396[a]	,000	0	.
M2	2,695[b]	.	28	.
M6	5,417[b]	1,021	28	1,000
M8	98,412[b]	94,016	28	,000
M9	81,536[b]	77,140	28	,000
M12	1,152E2	110,847	28	,000
M13	9,428[b]	5,032	28	1,000
M15	7,128[b]	2,732	28	1,000
M16	82,653[b]	78,257	28	,000
M17	4,907[b]	,511	28	1,000
M21	1,363E2	131,862	28	,000
M22	47,130[b]	42,734	28	,037
Boca	,000[b]	.	84	.

El estadístico de chi-cuadrado es la diferencia en las -2 log verosimilitudes entre el modelo final y el modelo reducido. El modelo reducido se forma omitiendo un efecto del modelo final. La hipótesis nula es que todos los parámetros de ese efecto son 0.

a. Este modelo reducido es equivalente al modelo final ya que la omisión del efecto no incrementa los grados de libertad.

b. Se han encontrado singularidades inesperadas en la matriz Hessiana. Esto indica que se deberán excluir algunas variables predictoras o que se deberán fusionar algunas categorías.

Se procede eliminando de una en una las variables no significativas hasta que solo queden variables significativas o hasta que la eliminación de alguna variable origine que el modelo no acierte correctamente el 100% de los casos. En el modelo nos quedamos con las variables M8, M12, M13, M15, M16, M17 y la posición de la boca.

CUADRO XI.5. *(Continuación)*[SPSS]

En el modelo que se obtiene se observa que todas las variables son significativas y se predice correctamente el 100% de los casos.

Contrastes de la razón de verosimilitud

Efecto	Criterio de ajuste del modelo	Contrastes de la razón de verosimilitud		
	-2 log verosimilitud del modelo reducido	Chi-cuadrado	gl	Sig.
Intersección	,516[a]	,000	0	.
M8	7,222E2	721,706	28	,000
M12	4,298E2	429,258	28	,000
M13	3,566E2	356,084	28	,000
M15	3,426E2	342,078	28	,000
M16	6,541E2	653,537	28	,000
M17	1,124E2	111,882	28	,000
Boca	,000[b]	.	84	.

El estadístico de chi-cuadrado es la diferencia en las -2 log verosimilitudes entre el modelo final y el modelo reducido. El modelo reducido se forma omitiendo un efecto del modelo final. La hipótesis nula es que todos los parámetros de ese efecto son 0.

a. Este modelo reducido es equivalente al modelo final ya que la omisión del efecto no incrementa los grados de libertad.

b. Se han encontrado singularidades inesperadas en la matriz Hessiana. Esto indica que se deberán excluir algunas variables predictoras o que se deberán fusionar algunas categorías.

Paso 13. En la tabla «Información del ajuste del modelo» podemos ver el grado de significación del modelo cuando se compara con el que no tiene ninguna variable independiente (solo la constante). Lo hace a partir de un estadístico Chi-cuadrado, determinando la probabilidad de obtener un Chi-cuadrado de ese valor o mayor si los coeficientes fuesen nulos. En este ejemplo, un $p < 0{,}001$ indica que el modelo final es significativamente distinto del que solo tiene la constante.

Información del ajuste del modelo

Modelo	Criterio de ajuste del modelo	Contrastes de la razón de verosimilitud		
	-2 log verosimilitud	Chi-cuadrado	gl	Sig.
Sólo la intersección	1,848E3			
Final	,516	1,848E3	252	,000

CUADRO XI.5. *(Continuación)*[SPSS]

Paso 14. En la tabla «Bondad de ajuste» realiza dos Chi-cuadrado diferentes con las desviaciones entre los valores observados y la predicción. Como $p > 0,05$ las desviaciones son pequeñas, es decir, el modelo se ajusta a los datos.

Bondad de ajuste

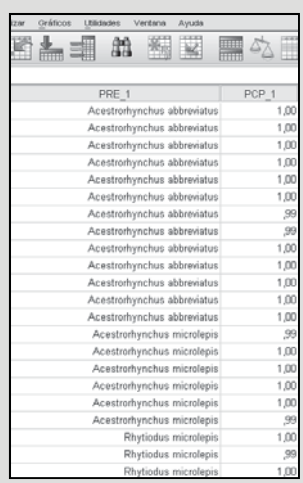
	Chi-cuadrado	gl	Sig.
Pearson	,259	7532	1,000
Desviación	,516	7532	1,000

Paso 15. Los valores de «Pseudo R-cuadrado» son altos indicando que un alto porcentaje de la variabilidad de los datos está explicada por el modelo. Según el r^2 de Nagelkerke explica un 100% de la variabilidad.

Pseudo R-cuadrado

Cox y Snell	,999
Nagelkerke	1,000
McFadden	1,000

Paso 16. En la matriz de datos, las últimas columnas son el grupo pronosticado y la probabilidad de pertenencia al grupo. Si se introduce una nueva fila o filas con los datos biométricos de uno o varios individuos, sin especificar a que especie pertenece, al correr de nuevo el análisis, se obtiene información del grupo de pertenencia y la probabilidad.



PRE_1	PCP_1
Acestrorhynchus abbreviatus	1,00
Acestrorhynchus abbreviatus	1,00
Acestrorhynchus abbreviatus	1,00
Acestrorhynchus abbreviatus	1,00
Acestrorhynchus abbreviatus	1,00
Acestrorhynchus abbreviatus	1,00
Acestrorhynchus abbreviatus	,99
Acestrorhynchus abbreviatus	,99
Acestrorhynchus abbreviatus	1,00
Acestrorhynchus abbreviatus	1,00
Acestrorhynchus abbreviatus	1,00
Acestrorhynchus abbreviatus	1,00
Acestrorhynchus abbreviatus	1,00
Acestrorhynchus microlepis	,99
Acestrorhynchus microlepis	1,00
Acestrorhynchus microlepis	1,00
Acestrorhynchus microlepis	1,00
Acestrorhynchus microlepis	1,00
Acestrorhynchus microlepis	,99
Rhytiodus microlepis	1,00
Rhytiodus microlepis	,99
Rhytiodus microlepis	1,00

Paso 17. El programa estima una función polinómica para cada clase (especie de pez) excepto para la última especie (*Pimelodus blochii*), que es la categoría de referencia. En la tabla «Estimaciones de los parámetros« (columna B) se muestran las estimaciones de todos los parámetros o coeficientes que multiplican a las variables independientes, incluidas las que crea el programa para cada nivel (excepto uno) de las variables categóricas. Estos son los coeficientes del modelo de regresión.

CUADRO XI.5. *(Continuación)*[SPSS]

Estimaciones de los parámetros

Especie[a]		B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
								Límite inferior	Límite superior
Acestrorhynchus abbreviatus	Intersección	-82,774	175320,031	,000	1	1,000			
	M8	246,189	5079,545	,002	1	,961	8,290E106	,000	.[b]
	M12	281,710	17596,695	,000	1	,987	2,214E122	,000	.[b]
	M13	-119,378	11991,708	,000	1	,992	1,428E-52	,000	.[b]
	M15	254,082	6133,518	,002	1	,967	2,220E110	,000	.[b]
	M16	-180,219	12752,214	,000	1	,989	5,394E-79	,000	.[b]
	M17	-45,705	3906,514	,000	1	,991	1,415E-20	,000	.[b]
	[Boca=1,00]	-8,443	175375,637	,000	1	1,000	,000	,000	.[b]
	[Boca=2,00]	8,995	175314,761	,000	1	1,000	8062,004	,000	.[b]
	[Boca=3,00]	-184,205	175279,227	,000	1	,999	1,002E-80	,000	.[b]
	[Boca=4,00]	0[c]	.	.	0	.	.	.	.
Acestrorhynchus microlepis	Intersección	-18,099	87359,591	,000	1	1,000			
	M8	-185,035	5108,126	,001	1	,971	4,368E-81	,000	.[b]
	M12	555,766	17520,963	,001	1	,975	2,324E241	,000	.[b]
	M13	-233,324	11998,502	,000	1	,984	4,663E-102	,000	.[b]
	M15	183,208	6180,664	,001	1	,976	3,682E79	,000	.[b]
	M16	-707,020	12871,037	,003	1	,956	8,813E-308	,000	.[b]
	M17	498,334	3999,765	,016	1	,901	2,653E216	,000	.[b]
	[Boca=1,00]	72,899	87483,179	,000	1	,999	4,567E31	,000	.[b]
	[Boca=2,00]	80,002	87348,676	,000	1	,999	5,553E34	,000	.[b]
	[Boca=3,00]	-25,108	87271,447	,000	1	1,000	1,246E-11	,000	.[b]
	[Boca=4,00]	0[c]	.	.	0	.	.	.	.
Rhytiodus microlepis	Intersección	96,155	118581,731	,000	1	,999			
	M8	-685,260	5265,173	,017	1	,896	2,484E-298	,000	.[b]
	M12	1031,237	17463,103	,003	1	,953	.[b]	,000	.[b]
	M13	-90,196	11994,847	,000	1	,994	6,734E-40	,000	.[b]
	M15	-86,295	6173,109	,000	1	,989	3,331E-38	,000	.[b]
	M16	22,381	12652,259	,000	1	,999	5,247E9	,000	.[b]
	M17	47,983	3934,860	,000	1	,990	6,898E20	,000	.[b]
	[Boca=1,00]	28,218	118661,419	,000	1	1,000	1,798E12	,000	.[b]
	[Boca=2,00]	-25,793	118573,425	,000	1	1,000	6,284E-12	,000	.[b]
	[Boca=3,00]	-318,651	118508,782	,000	1	,998	4,088E-139	,000	.[b]
	[Boca=4,00]	0[c]	.	.	0	.	.	.	.
Schizodon fasciatus	Intersección	-85,841	206357,907	,000	1	1,000			
	M8	150,872	5011,598	,001	1	,976	3,333E65	,000	.[b]
	M12	-366,752	17428,126	,000	1	,983	5,268E-160	,000	.[b]
	M13	485,599	11980,196	,002	1	,968	7,817E210	,000	.[b]
	M15	-191,027	6093,378	,001	1	,975	1,092E-83	,000	.[b]
	M16	-72,844	12613,732	,000	1	,995	2,313E-32	,000	.[b]
	M17	81,309	3813,509	,000	1	,983	2,051E35	,000	.[b]
	[Boca=1,00]	-4,444	206416,021	,000	1	1,000	,012	,000	.[b]
	[Boca=2,00]	-58,584	206353,931	,000	1	1,000	3,610E-26	,000	.[b]
	[Boca=3,00]	-87,828	206344,827	,000	1	1,000	7,189E-39	,000	.[b]
	[Boca=4,00]	0[c]	.	.	0	.	.	.	.

CUADRO XI.5. *(Continuación)*[SPSS]

Paso 18. Para estimar a que especie pertenece un individuo, con los coeficientes mostrados en la tabla anterior (Columna B), se calculan los valores *Z* para cada especie.

A modo de ejemplo se estima la probabilidad para el primer individuo de la matriz de datos y sólo se muestra como sería el cálculo de *Z* para dos especies: *Acestrorhynchus abbreviatus* y *Roeboides myersii*, realizando de la misma forma el cálculo de *Z* para el resto de especies.

Acestrorhynchus abbreviatus

$Z_{A.\ abbreviatus}$ = -82,774+ 246,189**M8* + 281,710**M12* - 119,378**M13* + 254,082**M15* - 180,219**M16* - 45,705**M17* - 8,443**Boca1* + 8,995**Boca2* - 184,205**Boca3*

$Z_{A.\ abbreviatus}$ = -82,774+ 246,189**0,3* + 281,710**0,1* - 119,378**0,23* + 254,082**0,27* - 180,219**0,17* - 45,705**0,17* - 8,443**0* + 8,995*1 - 184,205*0 = 39,9

Roeboides myersii

$Z_{R.\ myersii}$ = -194,293 + 301,160**M8* + 689,294**M12* - 97,882**M13* + 185,635**M15* - 282,630**M16* + 212,055**M17* + 91,876**Boca1* + 15,126**Boca2* - 268,625**Boca3*

$Z_{R.\ myersii}$ = -194,293 + 301,160**0,3* + 689,294**0,1* - 97,882**0,23* + 185,635**0,27* - 282,630**0,17* + 212,055**0,17* + 91,876**1* + 15,126**0* - 268,625**0* = -19,4

Una vez que se obtienen todos los valores *Z* para cada una de las especies, se calcula la probabilidad de pertenencia a cada especie. A modo de ejemplo, sólo se muestra la probabilidad para las especies *Acestrorhynchus abbreviatus* y *Roeboides myersii*.

$$\Pi_{A.abbreviatus} = \frac{e^{39,9}}{1 + e^{39,9} + e^{-57,7} + e^{-38,3} + e^{-132,8} + e^{-62,5} .... + e^{-18,565}} = 1$$

$$\Pi_{R.myersii} = \frac{e^{-19,4}}{1 + e^{39,9} + e^{-57,7} + e^{-38,3} + e^{-132,8} + e^{-62,5} .... + e^{-18,565}} = 0$$

Por lo tanto, según el modelo de regresión logística obtenido, el primer individuo pertenece a la especie *Acestrorhynchus abbreviatus* con una probabilidad de 1.

Capítulo XII

SERIES TEMPORALES

Una serie temporal es un conjunto de observaciones de una variable tomadas a lo largo de intervalos regulares de tiempo, como el número de automóviles vendidos por un fabricante cada mes durante los últimos diez años. Las series temporales aparecen prácticamente en todos los campos de actividad. El interés de su análisis estadístico radica en el estudio del comportamiento de la serie, lo que permite explicar sus variaciones y, sobre todo, en la posibilidad de predecir valores futuros.

El principio fundamental que subyace al análisis de las series de tiempo es que los valores actuales (y futuros) de la serie dependen en cierta medida de los valores que ha tomado en el pasado, de modo que podríamos considerar que la serie tiene una cierta inercia que impide –o hace poco probables– cambios demasiado bruscos en períodos cortos de tiempo. Esa inercia puede ser utilizada para la predicción, admitiendo que en el futuro se comportará de modo similar a como lo hizo en el pasado. El éxito en el análisis y en la predicción de valores futuros dependerá sobre todo de la disponibilidad de una muestra amplia y fiable, y de la regularidad de la serie. Lamentablemente muchas series son dinámicas o cambiantes, y no se comportan de acuerdo con esquemas rígidos, de manera que el modelo estadístico utilizado es solamente una aproximación más o menos razonable al verdadero funcionamiento de la serie. En esos casos las predicciones deben hacerse únicamente a plazo corto, y con un grado de fiabilidad limitado.

En una serie temporal los estudios más frecuentes que se suelen hacer son la creación de un modelo y el análisis de los ciclos.

XII.1. Modelo ARIMA

Existen muchos modelos estadísticos aplicados a las series temporales, entre los cuales destaca por su potencia y eficacia el modelo ARIMA (Box & Jenkins, 1976).

A la hora de hacer un modelo se deben tener en cuenta los siguientes análisis: la autocorrelación, si la serie crece, decrece o es estacionaria, la estacionalidad y, por último, ver si hay variables que pueden estar relacionadas con la variable dependiente.

XII.1.1. AUTOCORRELACIÓN

Para saber algo más de la serie es útil estudiar la correlación o asociación entre los valores de la serie en cada período y los valores en los períodos anteriores, es decir, la relación entre la serie y ella misma (autocorrelación) desplazada uno o más períodos.

Se suele utilizar la función de autocorrelación simple (*ACF*), la cual muestra el conjunto de coeficientes de autocorrelación, e indica hasta que punto la variable depende de ella misma.

XII.1.2. SERIES ESTACIONARIAS O NO ESTACIONARIAS

El modelo ARIMA supone –como condición previa– que la serie es estacionaria, es decir que su comportamiento se mantiene sin cambios en el tiempo, ya que todo el planteamiento de esta metodología se basa en que la evolución en el futuro será similar a la que ha tenido en el pasado (inercia de la serie). En especial se debe comprobar que su valor medio permanece estable (no tiene tendencia creciente o decreciente) y su variabilidad es también estable, no cambia (no aumenta o disminuye). Si no se cumplen estas condiciones es necesario hacer algún tipo de transformación de la variable, generalmente sencilla, que conduzca a una serie estacionaria. Si no existe ninguna transformación adecuada, el modelo ARIMA no debe ser aplicado.

XII.1.3. ESTACIONALIDAD

Las series temporales pueden ser estacionales o no estacionales. Muchas series temporales contienen variaciones estacionales, o pautas de comportamiento que se repiten con regularidad: El número de turistas es sistemáticamente mayor en verano que en invierno, el volumen de compras en unos grandes almacenes es más alto durante los primeros días de cada mes, el número de automóviles que transitan por una calle es mayor entre las 19h y las 21h que en el intervalo 22h-24h. Por tanto, la componente estacional se define como aquellos movimientos de la serie que se repiten de forma periódica, siendo la periodicidad inferior al año. La estacionalidad puede estar referida a diferentes períodos de tiempo (meses, días, horas, etc.), y en algunos casos es sumamente regular. Para detectar pautas de estacionalidad es muy útil representar gráficamente la serie.

Los modelos ARIMA permiten incluir una componente estacional, que puede ser utilizada con fines de predicción, y que presenta la misma estructura (p, d, q) que la parte no estacional. Estos componentes (p, d, q) se explican más adelante, cuando se analice la construcción del modelo ARIMA.

En aquellas series temporales en las que existe estacionalidad es necesario corregir este efecto, para una mejor interpretación del modelo, y se recurre a la desestacionalización o corrección estacional. Para realizar esta operación es necesario aislar en primer lugar la componente estacional, lo que posibilitará su posterior eliminación.

XII.1.4. COVARIABLES

Es posible incluir en el modelo adicionalmente otras variables explicativas o covariables que suponemos que tienen relación causal con nuestra variable dependiente, o que al menos están estrechamente relacionadas con ella.

Con ello, además de identificar posibles variables que influyen en nuestra serie temporal, si la relación entre estas covariables y nuestra variable dependiente es significativa, se debería esperar que el modelo mejore, es decir, que sea más predictivo.

XII.1.5. CONSTRUCCIÓN DEL MODELO

En los modelos ARIMA los valores de la serie se obtienen mediante la combinación de tres componentes (*AR*, *I*, *MA*) o también llamados (p, d, q):

a) *Componente auto regresiva (p)*. Muestra la parte del valor de la variable que depende de los valores anteriores, de la historia de la serie. Se obtiene mediante un modelo de regresión lineal en el que la variable dependiente es la componente *AR* del período actual, y la variable independiente o explicativa el valor inmediatamente anterior o un cierto número "*p*" de valores anteriores de la misma serie. Por ejemplo, el valor *AR* del mes de junio depende del valor de la serie en el mes de mayo ($p = 1$), o de los valores de marzo, abril y mayo ($p = 3$). El valor de *p*, el orden de auto regresión, debe ser fijado para cada aplicación.
b) *Componente de integración (d)*. Una serie es estacionaria si su media, su variabilidad y su estructura de correlaciones no cambian con el tiempo. Si la serie no es estacionaria, sino que tiene un comportamiento global creciente o decreciente, existe un componente de integración que debe ser determinado en primer lugar. La eliminación de este componente permite considerar la serie como estacionaria, y debe ser integrado posteriormente mediante un proceso inverso al de su eliminación. Para el cálculo de la componente de integración se obtienen las diferencias entre valores consecutivos de la serie, repitiendo este procedimiento sobre las diferencias obtenidas el número de veces "*d*" (orden de integración) que sea necesario para conseguir una serie estacionaria.
c) *Componente de media móvil (q)*. En general las series no se comportan de modo exacto en relación con el modelo, sino que existe un elemento residual, que explica las diferencias entre el valor observado en cada período y el valor obtenido mediante el modelo auto regresivo. Este elemento representa la innovación en la serie, la parte que no responde a la inercia o al comportamiento histórico anterior, sino que es nueva en cada período. Analizando la serie podemos determinar estos residuos o términos de innovación en los períodos de tiempo ya transcurridos. La componente *MA* de cada período se obtiene como media de un cierto número "*q*" de estos residuos en los períodos anteriores, siendo *q* el orden de medias móviles.

En general, los valores de *p*, *d*, y *q* son pequeños (cero o uno). Los valores mayores de estos parámetros del modelo permiten un mejor ajuste a los datos de la muestra, pero pueden producir también un efecto de sobreajuste al introducir en el modelo aspectos propios de la muestra disponible, no representativos de la población de la que se ha obtenido sino particulares de la muestra, de manera que el modelo puede ser muy adecuado para la muestra pero poco apropiado para la población de la que procede. En general se prefiere un modelo simple, que puede ser más adecuado para los fines de predicción, especialmente si la muestra no es muy grande, o si la serie no es muy estacionaria en su comportamiento sino ligeramente cambiante.

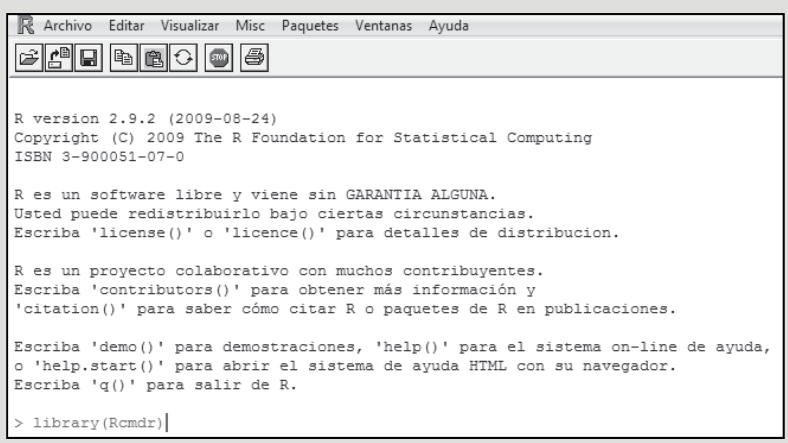
En el Cuadro XII.1 se dan varios ejemplos de cómo se construye un modelo ARIMA

Cuadro XII.1. Construcción de un modelo ARIMA

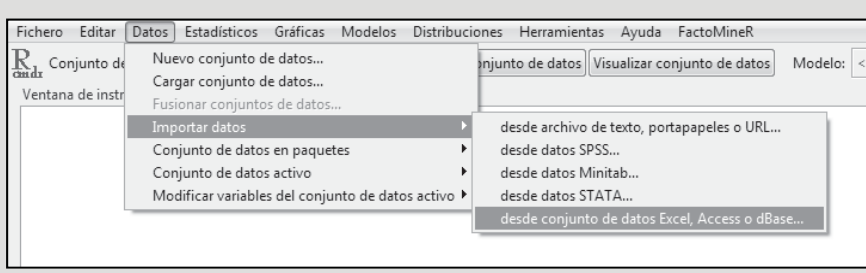
Aplicación con R.

EJEMPLO. La serie temporal que se va a analizar está en el archivo **Cuadro XII.1.xls**. Los datos muestran el factor de condición mensual (*FC*) de la anchoa (*Engraulis encrasicolus)* a lo largo de varios años. *FC* es un indicador del estado reproductivo de los individuos. El objetivo es determinar el efecto de los valores medios mensuales de la temperatura (en ºC), clorofila (en mg m^{-3}) y estabilidad de la columna de agua (en m^3s^{-3}) sobre el estado reproductivo de esta especie. Como *FC* y el resto de variables pueden covariar con el tiempo, es necesario cuantificar el posible efecto del tiempo sobre *FC*, en otras palabras, introducir el potencial efecto del tiempo sobre *FC* en el modelo, para poder determinar el efecto real de las variables mencionadas anteriormente sobre *FC*.

Paso 1. Iniciamos R y elegimos el directorio de trabajo. Son necesarios los paquetes "Rcmdr", "forecast", "nortest" y "seewave". Si los paquetes no están instalados, la instalación se realiza como se explica en el Apéndice II. Se introduce la instrucción «library(Rcmdr)».

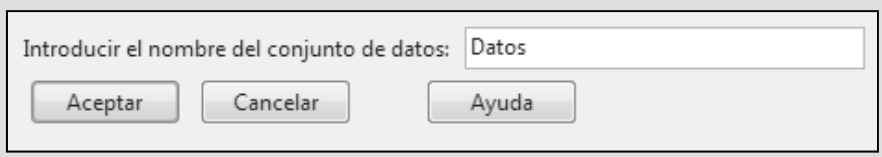

Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> library(Rcmdr)

Paso 2. Se selecciona «Datos» ⇨ «Importar datos» ⇨ «desde conjunto de datos Excel, Access o dBase».


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda FactoMineR
Nuevo conjunto de datos...
Cargar conjunto de datos...
Fusionar conjuntos de datos...
Importar datos
Conjunto de datos en paquetes
Conjunto de datos activo
Modificar variables del conjunto de datos activo
desde archivo de texto, portapapeles o URL...
desde datos SPSS...
desde datos Minitab...
desde datos STATA...
desde conjunto de datos Excel, Access o dBase...
Visualizar conjunto de datos
Modelo:

CUADRO XII.1. *(Continuación)*[R]

Paso 3. Aparece un pequeño cuadro de diálogo para introducir el nombre del conjunto de datos, en el que dejamos el nombre por defecto «Datos» y pulsamos «Aceptar». Posteriormente aparecen los archivos del directorio de trabajo, en el que elegimos **Cuadro XII.1.xls** y pulsamos «Abrir».

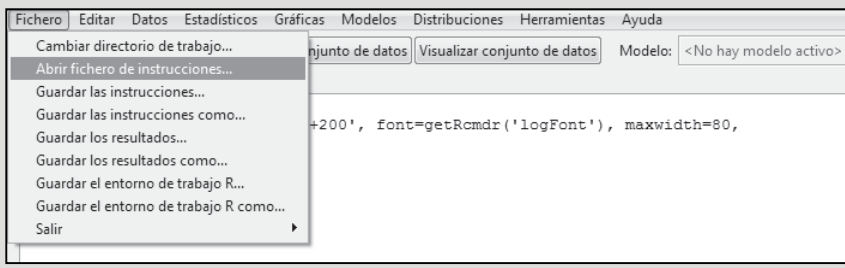


Paso 4. En R Commander pulsamos en «Visualizar conjunto de datos». Es importante inspeccionar los datos y comprobar que no existan líneas, lo cual a veces ocurre al final del archivo, que tengan celdas con el código <NA>. Si existen estas líneas, hay que eliminarlas.

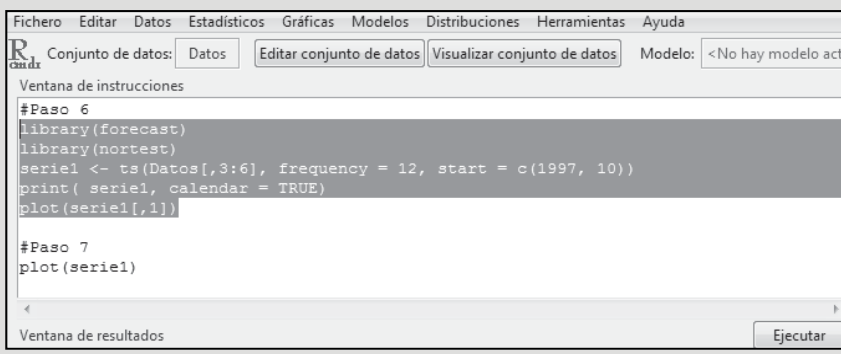
	Año	Mes	FC	Clorofila	Temperatura	Estabilidad
1	1997	10	0.00800	0.22	23.10	531.0
2	1997	11	0.00210	0.54	19.56	1110.0
3	1997	12	0.00195	0.56	16.88	1389.0
4	1998	1	0.00196	0.38	15.72	117.0
5	1998	2	0.00205	0.49	14.82	445.5
6	1998	3	0.00592	0.51	14.78	962.5
7	1998	4	0.01040	0.45	15.83	863.5
8	1998	5	0.01290	0.30	18.33	503.5
9	1998	6	0.02140	0.25	21.95	233.5
10	1998	7	0.03450	0.25	24.60	327.5
11	1998	8	0.04070	0.25	26.02	568.0
12	1998	9	0.02440	0.29	24.06	230.5
13	1998	10	0.01300	0.36	21.84	479.5
14	1998	11	0.00300	0.35	19.12	769.5
15	1998	12	0.00220	0.44	16.36	1100.0
16	1998	1	0.00196	0.53	15.15	1340.5
17	1999	2	0.00380	0.71	14.11	1267.5
18	1999	3	0.00500	0.62	14.27	2031.0
19	1999	4	0.02030	0.71	15.31	977.0
20	1999	5	0.03120	0.26	18.44	758.0
21	1999	6	0.02900	0.25	22.03	623.0
22	1999	7	0.03020	0.22	24.41	184.5
23	1999	8	0.02870	0.19	26.47	272.5
24	1999	9	0.00310	0.25	26.01	745.5
25	1999	10	0.00200	0.21	23.55	301.0
26	1999	11	0.00210	0.34	20.55	1119.0

CUADRO XII.1. *(Continuación)*[R]

Paso 5. A continuación se carga el archivo de instrucciones **Cuadro XII.1.R** seleccionando «Fichero» ⇨ «Abrir fichero de instrucciones…» y no se guarda la salida.



Paso 6. Se seleccionan las primeras líneas y se pulsa «Ejecutar». Los resultados y cada una de las instrucciones se explican a continuación.



Las primeras instrucciones cargan los paquetes necesarios.

```
library(forecast)
library(nortest)
```

Se construye la serie temporal. Esta operación asigna a los datos el formato adecuado en R para su tratamiento estadístico posterior. Se define la serie, que llamaremos «serie1», que incluye las columnas 3 a 6 de la matriz de datos. La periodicidad mensual (12) empieza en octubre (10) de 1997.

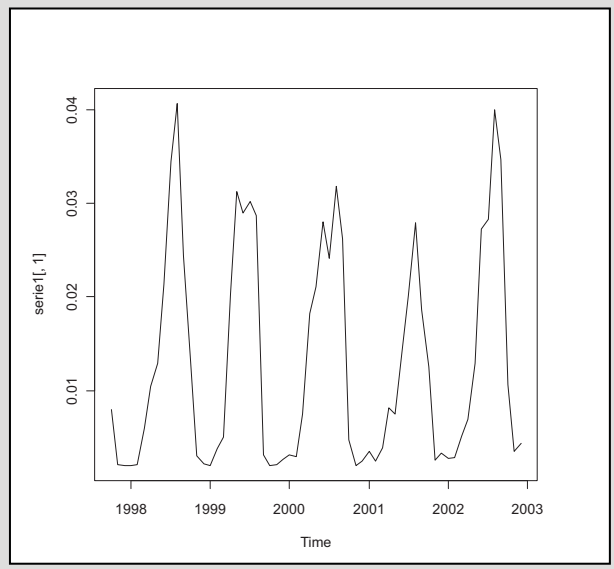
```
serie1 <- ts(Datos[,3:6], frequency = 12, start = c(1997, 10))
```

Imprimimos y representamos gráficamente la variable de interés, *FC*, la primera de la serie múltiple creada.

```
print( serie1, calendar = TRUE) # imprimimos la serie completa
plot(serie1[,1]) # gráfico de secuencia de la primera variable (FC), para la segunda variable sería [,2] y si sólo se hubiese capturado una variable no se pondría nada más que serie1, eliminando [,1]
```

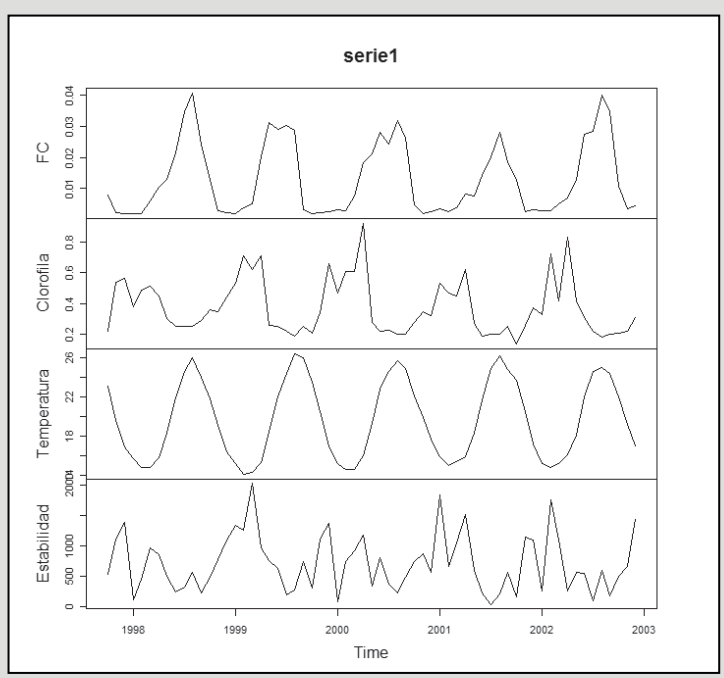
CUADRO XII.1. *(Continuación)*[R]

Se observa una clara estacionalidad anual, ya que todos los años se repite la misma pauta, con valores bajos en invierno y altos en verano. Por otra parte la serie en su conjunto no muestra tendencia (es horizontal), ni ciclo.



Paso 7. Se selecciona la siguiente instrucción que permite graficar todas las variables y se pulsa «Ejecutar».

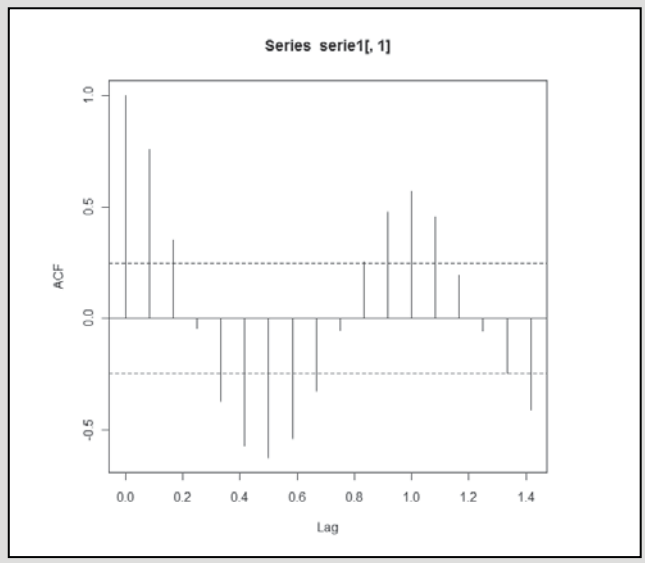
plot(serie1) *# gráfico de secuencia de todas las variables*



CUADRO XII.1. *(Continuación)*[R]

Paso 8. Autocorrelación. Función de autocorrelación simple y parcial. Se utilizan para la construcción del modelo ARIMA. Para que funcione si faltan datos en la serie, se añade la instrucción «na.action=na.pass».

acf(serie1[,1], na.action=na.pass) *# gráfico de la función de autocorrelación simple*

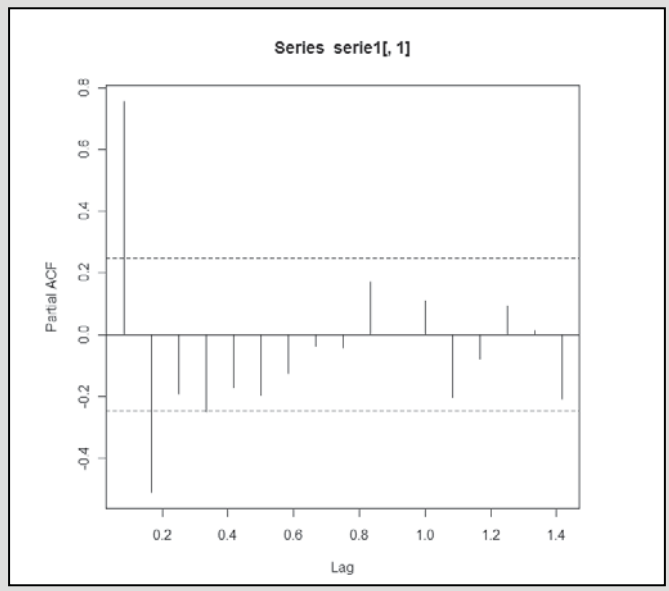


La función de autocorrelación simple (*ACF*), que se observa en el gráfico anterior, muestra el conjunto de coeficientes de autocorrelación, e indica hasta que punto la variable depende de ella misma. Cuando el valor de la serie Y_t depende fuertemente del valor anterior Y_{t-1} existirá un coeficiente de correlación alto R_1 entre la serie y la misma serie desplazada un período, pero dado que Y_{t-1} depende a su vez del valor anterior a él Y_{t-2}, también será alto el coeficiente de correlación con la serie desplazada dos períodos, R_2, y así sucesivamente. El gráfico muestra coeficientes que disminuyen gradualmente. R representa en esta función el coeficiente correspondiente al retardo cero, que siempre vale uno (no debe ser interpretado, solo sirve como referencia de comparación).

Es posible que una parte de esa dependencia que mide el coeficiente de correlación simple sea en realidad indirecta, a través de los valores intermedios, por lo que solamente los valores más próximos al actual deben ser considerados (tal vez uno solo). Por ello, el coeficiente de correlación parcial (*ACF* parcial o *ACFP*) es más adecuado para analizar estas relaciones de dependencia que el coeficiente de correlación simple, ya que al eliminar la influencia de las restantes variables (los valores intermedios) aísla el efecto propio de cada retardo sobre el valor actual de la serie, y permite estudiar la correlación entre la serie y la misma serie retardada sin incluir los efectos indirectos.

pacf(serie1[,1], na.action=na.pass) *# gráfico de la función de autocorrelación parcial*

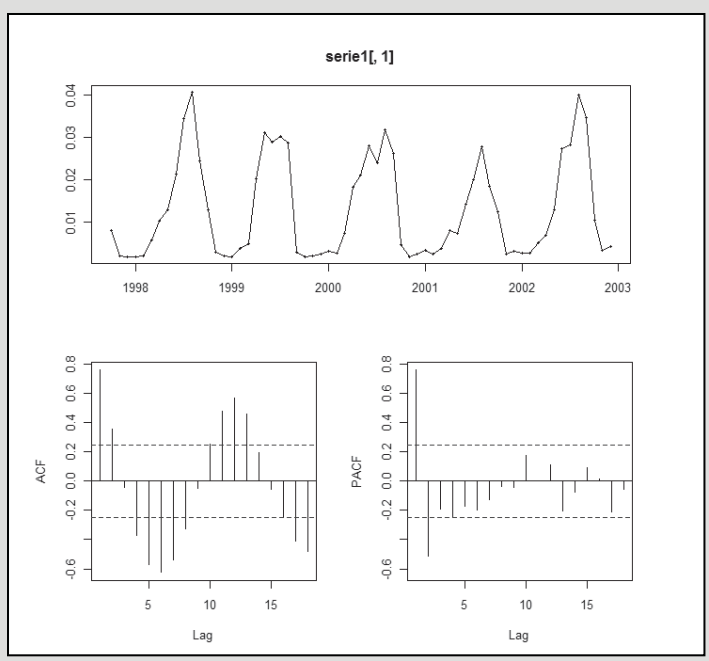
CUADRO XII.1. *(Continuación)*[R]



En la función de autocorrelación parcial comprobamos que solamente los dos primeros son significativos, y el primero tiene un valor elevado. Con estos resultados, un modelo adecuado podría tener $p = 1$ o $p = 2$ y $q = 0$.

La función tsdisplay de R representa en un solo paso el gráfico de secuencia, y ambas funciones de autocorrelación.

tsdisplay(serie1[,1])



CUADRO XII.1. *(Continuación)*[R]

Paso 9. Serie estacionaria o no estacionaria. Para encontrar un valor adecuado del parámetro "*d*" podemos utilizar la función «ndiffs», que devuelve el orden de diferenciación más conveniente. Como en los casos anteriores se selecciona y se pulsa «Ejecutar».

ndiffs(serie1[,1])

```
> ndiffs(serie1[,1])
[1] 0
```

En este caso *d* = 0, por lo tanto, la serie es estacionaria. Si la serie no fuera estacionaria sería suficiente con poner el valor de "*d*" recomendado, ya que lo que hace el modelo ARIMA es diferenciar la serie "*d*" veces, con lo que la serie se convierte en estacionaria, y una vez aplicado el modelo integra la serie resultante "*d*" veces para deshacer el cambio.

Paso 10. Construcción de un modelo Arima mediante un modelizador experto. R dispone de un procedimiento, «auto.arima», que busca entre los diferentes valores de los parámetros *p*, *q*, y *d*, tanto para la parte regular como para la parte estacional, el modelo más adecuado. Emplea el algoritmo de Hyndman & Khandakar (2008), que elige los parámetros minimizando el *AIC*, y los valores de *d* y *D* mediante test de raíz unitaria. Un método paso a paso (para lo cual debe incluirse la opción stepwise=TRUE) permite limitar el campo de búsqueda, reduciendo el esfuerzo de cálculo y el tiempo total necesario (en todo caso, el proceso puede tardar a veces varios minutos).

t1 <- auto.arima(serie1[,1],stepwise=TRUE) *# modelizador experto*
t1 *# imprime los resultados del modelo ARIMA elegido*

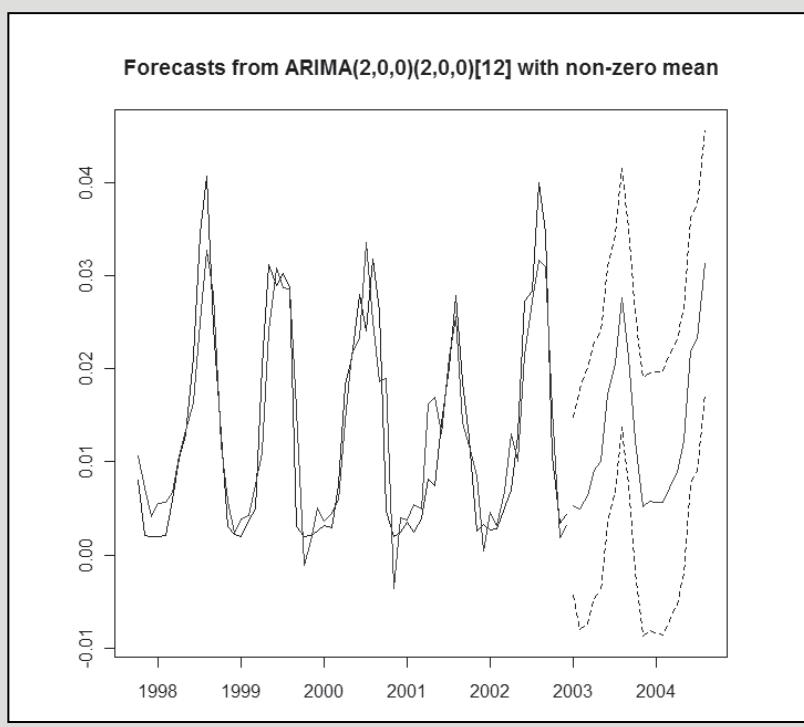
En nuestro ejemplo encuentra y ajusta un modelo ARIMA(2,0,0)(2,0,0), que considera óptimo.

```
> t1 <- auto.arima(serie1[,1],stepwise=TRUE)

> t1
Series: serie1[, 1]
ARIMA(2,0,0)(2,0,0)[12] with non-zero mean

Call: auto.arima(x = serie1[, 1], stepwise = TRUE)

Coefficients:
         ar1      ar2    sar1    sar2  intercept
      0.9185  -0.3629  0.2351  0.5494     0.0135
s.e.  0.1247   0.1192  0.0988  0.1115     0.0035

sigma^2 estimated as 2.380e-05:  log likelihood = 239.29
AIC = -466.59   AICc = -465.09   BIC = -453.73
```

CUADRO XII.1. *(Continuación)*[R]

Paso 11. Predicción de valores futuros del modelo Arima mediante modelizador experto. La función «forecast» de R calcula las predicciones para el número de períodos *h* deseado. La función «plot» representa gráficamente estos valores, que añade a los valores observados de la serie (a partir del último). También representa los intervalos de confianza para las predicciones. Se pueden colorear «shaded = TRUE» las bandas de confianza, o simplemente dibujar líneas. La instrucción «lines» a continuación de «plot» permite representar los valores ajustados, o predicciones del modelo para todos los períodos observados.

forecast(t1, h=20,level=c(95)) *# imprime los valores de predicción junto con el intervalo de confianza para cada valor*
plot(forecast(t1, h=20,level=c(95)), shaded = FALSE) *# representa serie original + predicciones*
lines(serie1[,1]-t1$residuals, col = "blue") *# superpone el valor ajustado = original -residuo*

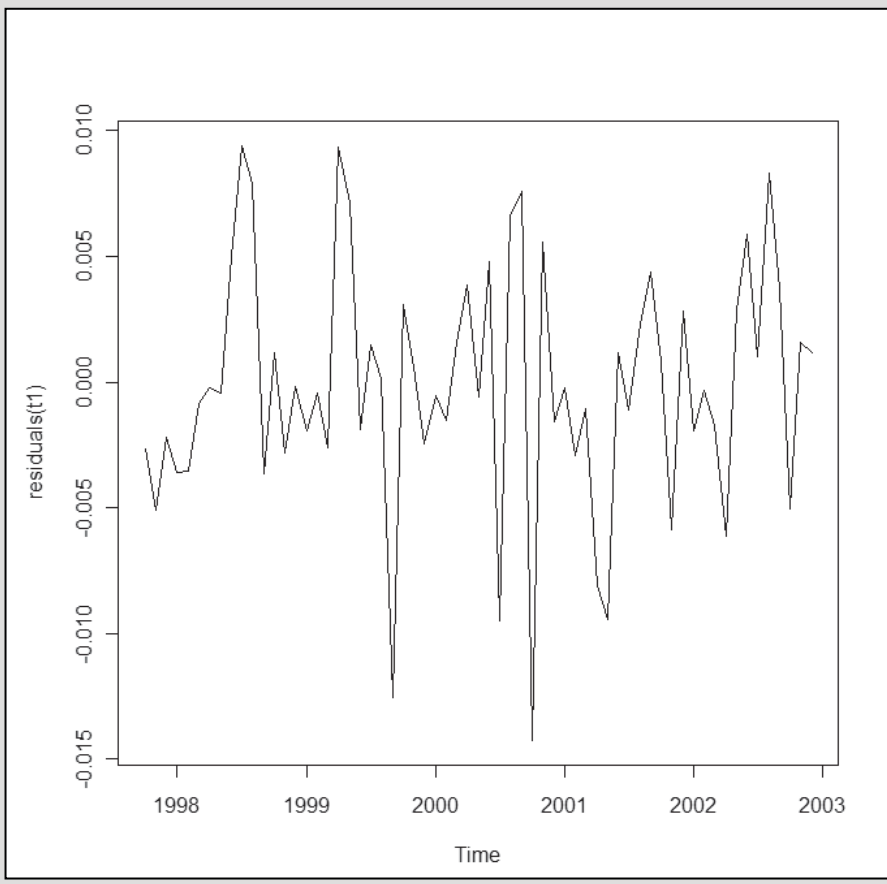



```
> forecast(t1, h=20,level=c(95))
         Point Forecast        Lo 95      Hi 95
Jan 2003    0.005247991 -0.004313171 0.01480915
Feb 2003    0.004956019 -0.008025986 0.01793802
Mar 2003    0.006335411 -0.007436090 0.02010691
Apr 2003    0.009086000 -0.004724284 0.02289628
May 2003    0.010097839 -0.003731092 0.02392677
Jun 2003    0.017125859  0.003258272 0.03099345
```

CUADRO XII.1. *(Continuación)*[R]

Paso 12. Diagnósticos del modelo Arima mediante modelizador experto. Es conveniente comprobar que el modelo elegido es completo y no deben añadirse nuevos elementos. Para ello comenzamos analizando los residuos con un gráfico.

```
plot(residuals(t1))
```

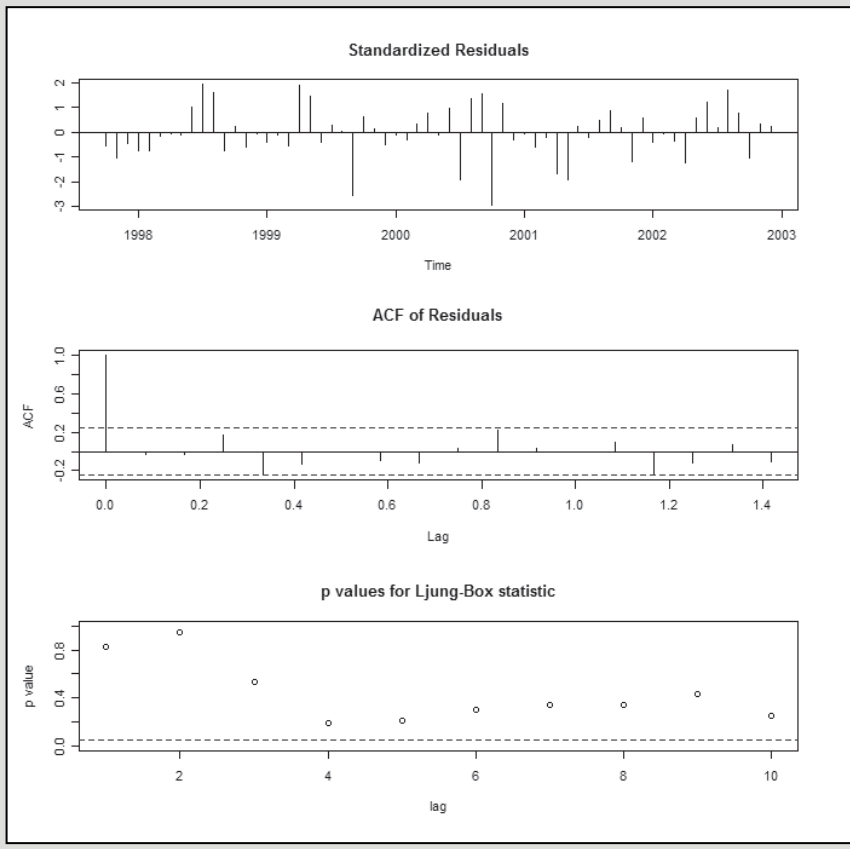


Podemos comprobar como los residuos parecen comportarse aleatoriamente, sin secuencias predecibles o una estructura de autocorrelaciones, es decir de acuerdo con el modelo conocido como "ruido blanco".

La función «tsdiag» elabora gráficos de residuos, *FAS* de residuos, y valores *p* de Ljung-Box para cada retardo.

```
tsdiag(t1)
```

CUADRO XII.1. *(Continuación)*[R]



En el gráfico de residuos tipificados (similar, salvo cambio de escala, al gráfico anterior) podemos comprobar como la mayoría son pequeños. El gráfico *ACF* no debe mostrar autocorrelaciones importantes (recordemos que la primera, siempre igual a 1, no cuenta), como en este ejemplo. Al observar los *p*-valores correspondientes a la prueba de Ljung-Box, para todos los retardos no se rechaza la hipótesis de que las autocorrelaciones estimadas son normales con media cero y varianza 1/n. Todos o casi todos los *p* valores deben estar por encima de la línea azul punteada del 5%. Por lo tanto, parece razonable concluir que los residuos se comportan como un ruido blanco, y el modelo es adecuado.

Existe otra función en R que realiza un test conjunto para contrastar la hipótesis nula de que se trata de ruido blanco:

Box.test(t1$resid, 12, type="Ljung")

```
> Box.test(t1$resid, 12, type="Ljung")

        Box-Ljung test

data:  t1$resid
X-squared = 12.6802, df = 12, p-value = 0.3927
```

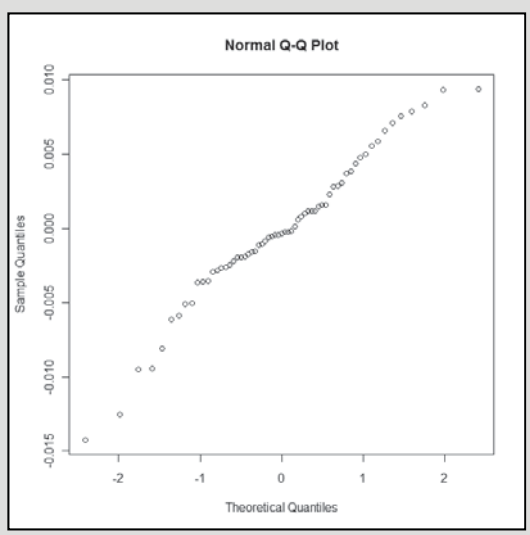
CUADRO XII.1. *(Continuación)*[R]

El valor p = 0,3927 mayor que el nivel de significación habitual 0,05 indica que se puede aceptar que se trata de ruido blanco y, por lo tanto, los residuos no conservan una estructura que deba ser explicada mediante un modelo más complejo.

Podemos comprobar también la distribución Normal de los residuos mediante un gráfico QQ utilizando la función «qqnorm» y aplicar el test de Shapiro-Wilk y el de Kolmogorov-Smirnov con la corrección de Lilliefors a los residuos.

```
qqnorm(t1$resid)
Shapiro<-shapiro.test(residuals(t1))
Shapiro
Lillie<-lillie.test(residuals(t1))
Lillie
```

El gráfico debe mostrar la serie de puntos siguiendo aproximadamente una línea recta. Una pequeña desviación en los extremos es aceptable. Los residuos tienen una distribución Normal, ya que tanto en el test de Shapiro-Wilk como en el de Kolmogorov-Smirnov $p > 0,05$.



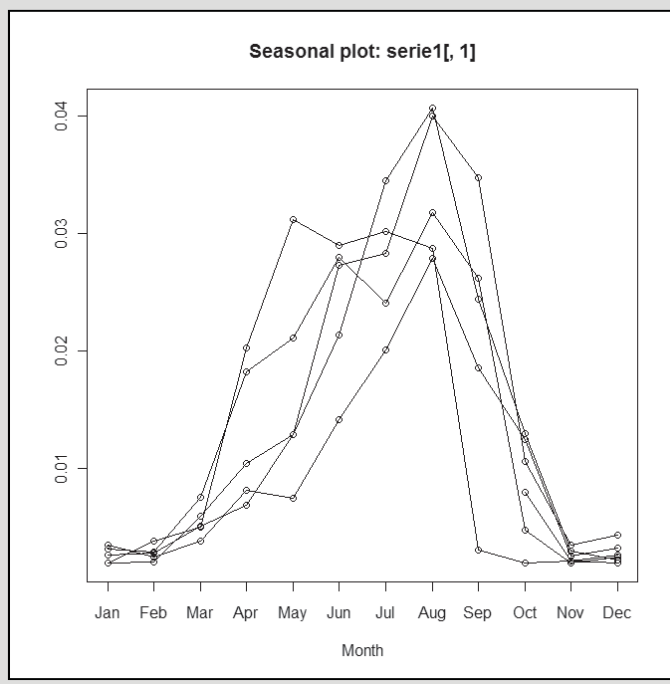

```
> Shapiro

        Shapiro-Wilk normality test

data:  residuals(t1)
W = 0.9715, p-value = 0.1513


> Lillie<-lillie.test(residuals(t1))

> Lillie

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuals(t1)
D = 0.0952, p-value = 0.1673
```

CUADRO XII.1. *(Continuación)*[R]

Paso 13. Desestacionalización. En las series con estacionalidad acusada, como la de nuestro ejemplo, es difícil apreciar las variaciones que se ocultan o enmascaran detrás de la estacionalidad y, por lo tanto, averiguar los cambios que se producen. Para ello podemos analizar con más detalle las variaciones estacionales y eliminar la estacionalidad de la serie para estudiar el resto.

a) Valores de estacionalidad. Se pueden observar con un gráfico adecuado, como el que produce la función «seasonplot» de R.

```
seasonplot(serie1[,1])
```



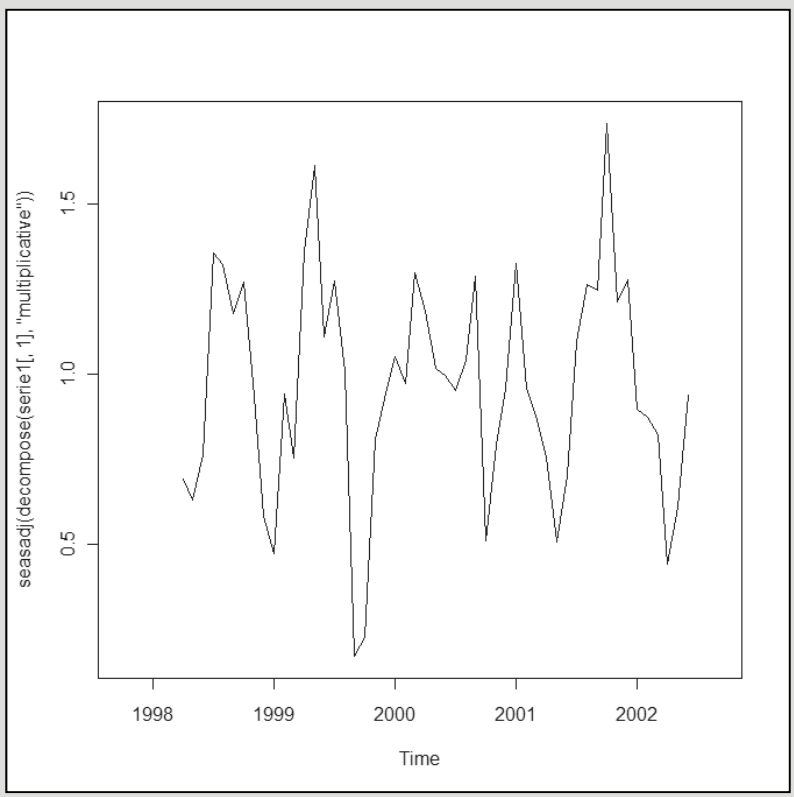
Como se puede apreciar claramente, los valores de estacionalidad alta son los que corresponden al mes de agosto y próximos a este mes, y la estacionalidad baja ocurre en los meses que van de noviembre a febrero, todos ellos en niveles similares.

b) Serie desestacionalizada. Es fácil eliminar la estacionalidad con la función «seasadj», que produce un ajuste estacional. Cuando la serie no está completa, en lugar de la serie original se utiliza una serie modificada en la que los valores perdidos se sustituyen por interpolación de valores próximos, mediante la función «na.approx». El resultado es aproximado, pero razonablemente válido a efectos de representación gráfica de la serie desestacionalizada.

```
s3 <- na.approx(serie1[,1])
plot(seasadj(decompose(s3,"multiplicative")))
```

CUADRO XII.1. *(Continuación)*[R]

El gráfico que se muestra a continuación permite observar los cambios en la serie descontando el efecto de la estacionalidad.

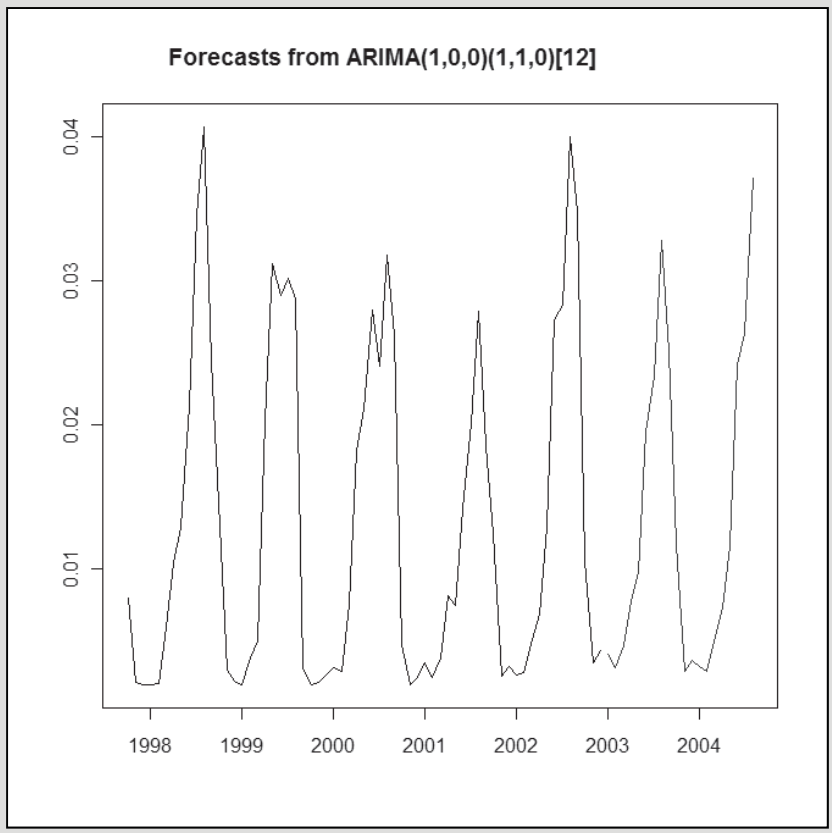


Paso 14. Construcción de un modelo Arima previamente especificado. En lugar de utilizar el modelizador experto, podemos decidir –a partir de los gráficos de autocorrelación o del conocimiento de la serie– un modelo concreto, por ejemplo con valores más sencillos para los parámetros, como ARIMA(1,0,0)(1,1,0). Para ello utilizamos la función «arima», de R.

```
t1 <- arima(serie1[,1], order = c(1,0,0), seasonal = list(order=c(1,1,0)))
t1
plot(forecast(t1, h=20), plot.conf = FALSE) # plot.conf =TRUE dibuja intervalos de confianza
```

El gráfico muestra como las predicciones, en color azul, siguen la misma secuencia observada en los datos disponibles. A menudo es preferible un modelo más sencillo cuando las predicciones son razonables.

CUADRO XII.1. *(Continuación)*[R]



Para ver si es mejor modelo que el obtenido con el sistema experto, la varianza estimada de los residuos «sigma^2 estimated» debe ser pequeña, como así lo es en este caso ($2.823*10^{-5}$), y el valor de *AIC* es mejor cuanto más pequeño (o más negativo). En el caso del modelo especificado *AIC* es -377,7, mayor que en el modelo del sistema experto (-466,59, véase paso 10). Por lo tanto, el modelo obtenido por el sistema experto es mejor.

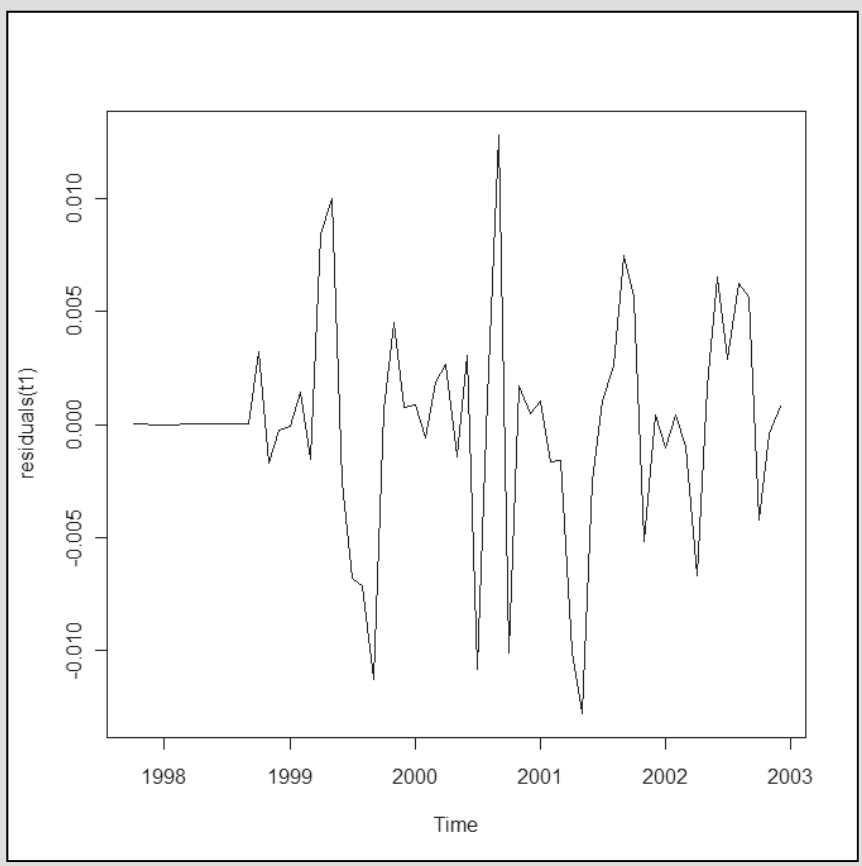
```
> t1
Series: serie1[, 1]
ARIMA(1,0,0)(1,1,0)[12]

Call: arima(x = serie1[, 1], order = c(1, 0, 0), seasonal = list(order = c(1, 1, 0)))

Coefficients:
         ar1     sar1
      0.6051  -0.5995
s.e.  0.1088   0.1121

sigma^2 estimated as 2.823e-05:  log likelihood = 191.85
AIC = -377.7   AICc = -377.19   BIC = -371.9
```

CUADRO XII.1. *(Continuación)*[R]

Paso 15. Diagnósticos del modelo Arima previamente especificado. Como se hizo anteriormente con el modelo experto, se puede comprobar si el modelo especificado es completo y no deben añadirse nuevos elementos. Para ello comenzamos analizando los residuos con un gráfico.

```
plot(residuals(t1))
```

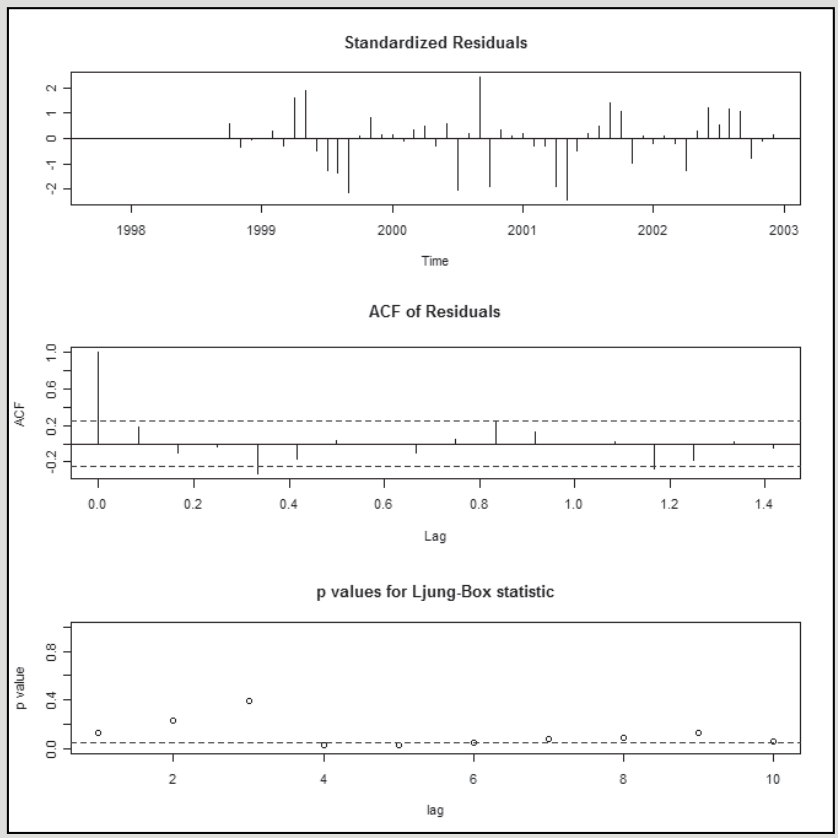
Al igual que en el modelo experto, los residuos parecen comportarse aleatoriamente, sin secuencias predecibles o una estructura de autocorrelaciones.



La función «tsdiag» elabora gráficos de residuos, *FAS* de residuos, y valores *p* de Ljung-Box para cada retardo.

```
tsdiag(t1)
```

CUADRO XII.1. *(Continuación)*[R]



En el gráfico de residuos tipificados se observa como la mayoría son pequeños. El gráfico *ACF* no debe mostrar autocorrelaciones importantes, como en este ejemplo. Al observar los *p*-valores correspondientes a la prueba de Ljung-Box se observa por qué este modelo es peor, ya que algunos *p* valores están por debajo de la línea azul punteada del 5%.

Al contrastar la hipótesis nula de que se trata de ruido blanco:

Box.test(t1$resid, 12, type="Ljung")

El valor p = 0,083 mayor que el nivel de significación habitual 0,05 indica que se puede aceptar que se trata de ruido blanco.

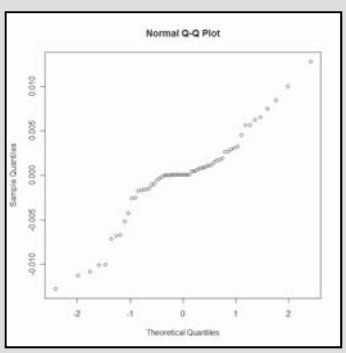
```
> Box.test(t1$resid, 12, type="Ljung")

        Box-Ljung test

data:  t1$resid
X-squared = 19.2327, df = 12, p-value = 0.08307
```

CUADRO XII.1. *(Continuación)*[R]

Por último, comprobamos la distribución Normal de los residuos:

```
qqnorm(t1$resid)
Shapiro<-shapiro.test(residuals(t1))
Shapiro
Lillie<-lillie.test(residuals(t1))
Lillie
```

El gráfico muestra que la serie de puntos no sigue una línea tan recta como en el modelo experto y, por tanto, de nuevo se observa que el modelo es peor. Los residuos no tienen una distribución Normal, ya que tanto en el test de Shapiro-Wilk como en el de Kolmogorov-Smirnov $p < 0,05$.




```
> Shapiro

        Shapiro-Wilk normality test

data:  residuals(t1)
W = 0.9241, p-value = 0.0008228


> Lillie<-lillie.test(residuals(t1))

> Lillie

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuals(t1)
D = 0.1786, p-value = 3.199e-05
```

Paso 16. Introducción de covariables. Es posible incluir en el modelo nuevas variables explicativas. Para ello solo tenemos que añadir la opción «xreg» con las variables independientes cuyo efecto queremos medir, que en este ejemplo ocupan las columnas 2 a 4 de la serie temporal (la primera es *FC*).

```
t1 <- arima(serie1[,1], order = c(2,0,0), seasonal = list(order=c(2,0,0)),
xreg =serie1[,2:4])
t1
```

```
> t1
Series: serie1[, 1]
ARIMA(2,0,0)(2,0,0)[12] with non-zero mean

Call: arima(x = serie1[, 1], order = c(2, 0, 0), seasonal = list(order = c(2, 0, 0)),      xreg = serie1[, 2:4])

Coefficients:
         ar1      ar2    sar1    sar2  intercept  Clorofila  Temperatura  Estabilidad
      0.8627  -0.3097  0.1802  0.4780    -0.0290     0.0017       0.0021            0
s.e.  0.1198   0.1189  0.1065  0.1286     0.0075     0.0055       0.0007            0

sigma^2 estimated as 2.240e-05:  log likelihood = 243.6
AIC = -469.19   AICc = -465.8   BIC = -449.9
```

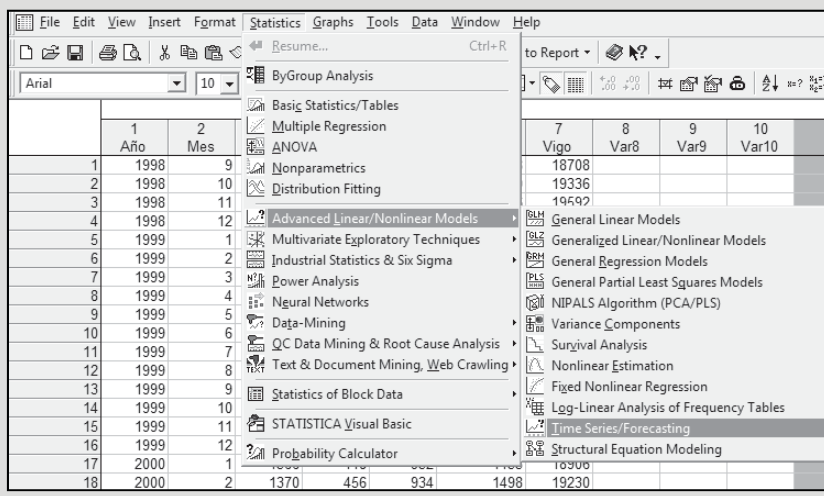
El valor de *AIC* con las covariables es -469,19, mientras que sin ellas era (-466,59, véase paso 10). La diferencia es tan pequeña que podemos afirmar que las covariables no añaden información útil y, por lo tanto, el estado reproductivo *FC* no depende de la temperatura, clorofila, ni estabilidad de la columna de agua.

Cuadro XII.1. *(Continuación)*[STATISTICA]

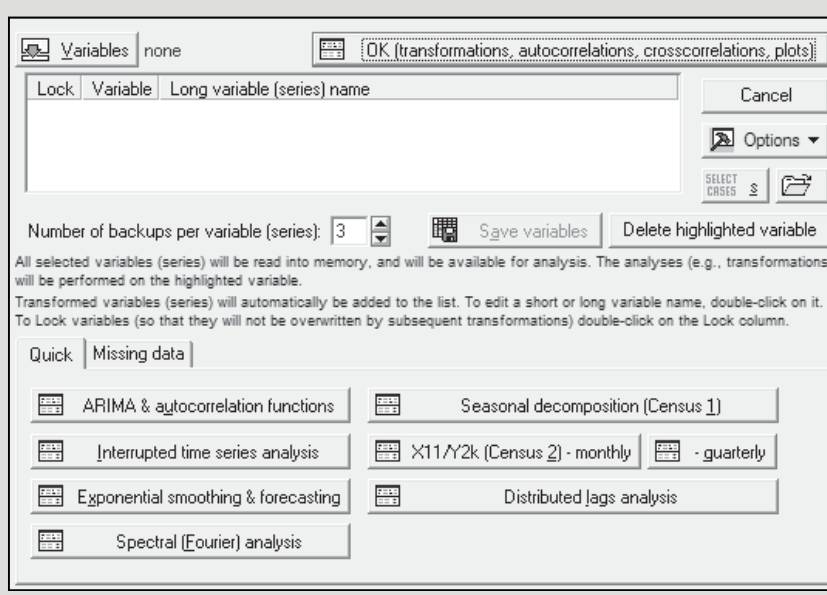
Aplicación con STATISTICA.

EJEMPLO. El archivo **Cuadro XII.1.sta** tiene datos del número de desempleados en diferentes ciudades. El objetivo es analizar la serie temporal de la ciudad de Vigo y determinar si el número de desempleados en las otras ciudades contribuye a mejorar la predicción del modelo.

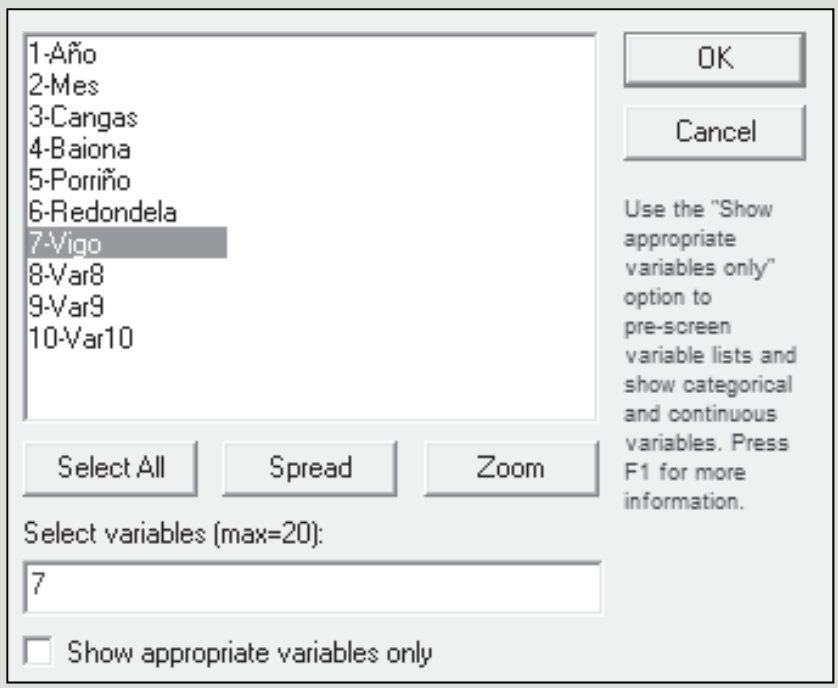
Paso 1. En primer lugar se accede a «Statistics» ⇨ «Advanced/Nonlinear Models» ⇨ «Time Series/Forecasting».



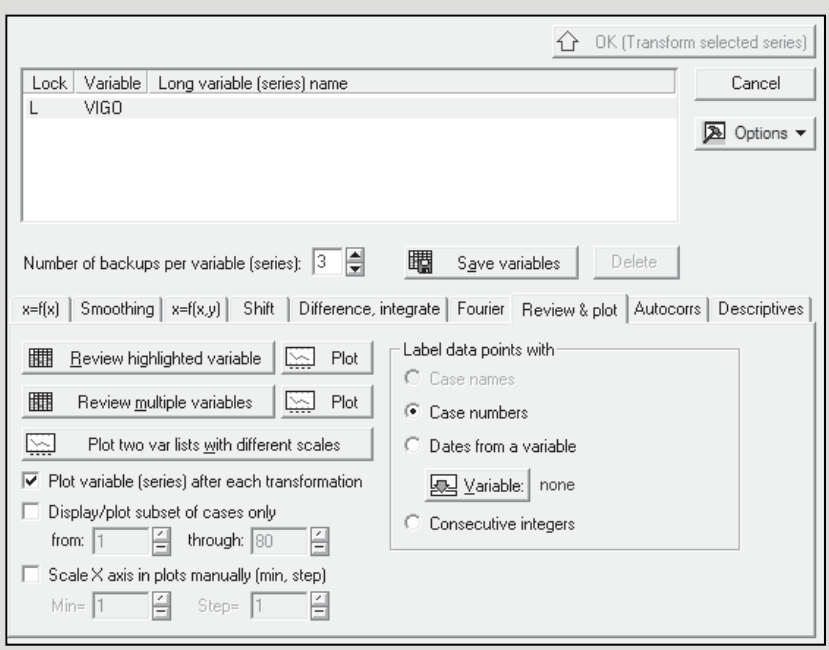
Paso 2. A continuación se pulsa en «Variables» y en la pantalla que aparece se selecciona la variable «Vigo».


Variables none
OK (transformations, autocorrelations, crosscorrelations, plots)
Lock Variable Long variable (series) name
Cancel
Options
SELECT CASES
Number of backups per variable (series): 3
Save variables
Delete highlighted variable
All selected variables (series) will be read into memory, and will be available for analysis. The analyses (e.g., transformations) will be performed on the highlighted variable.
Transformed variables (series) will automatically be added to the list. To edit a short or long variable name, double-click on it. To Lock variables (so that they will not be overwritten by subsequent transformations) double-click on the Lock column.
Quick Missing data
ARIMA & autocorrelation functions
Seasonal decomposition (Census 1)
Interrupted time series analysis
X11/Y2k (Census 2) - monthly
- quarterly
Exponential smoothing & forecasting
Distributed lags analysis
Spectral (Fourier) analysis

Cuadro XII.1. *(Continuación)* STATISTICA

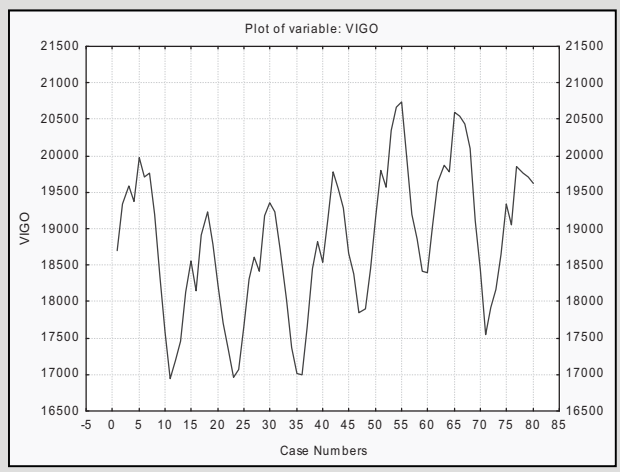


Paso 3. Para obtener una representación de los datos, en la ventana del paso 2, se pulsa en «OK (transformations, autocorrelations, crosscorrelations, plots)» y en la pestaña «Review & plot», de la ventana que aparece, pulsamos en «Plot» de «Review highlighted variable».

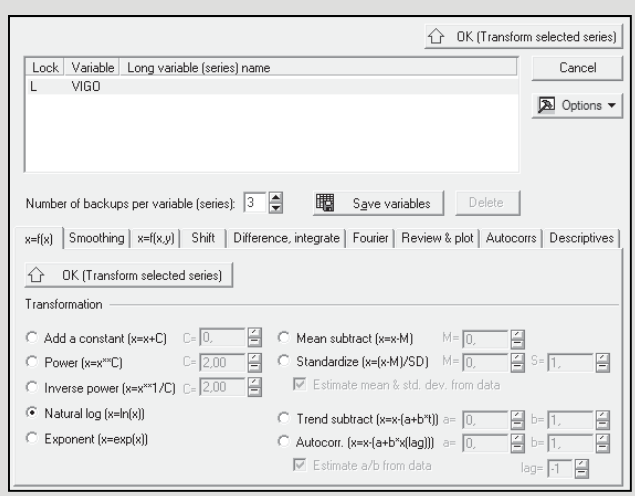


Cuadro XII.1. *(Continuación)*[STATISTICA]

Paso 4. La serie muestra una clara componente estacional, con picos máximos en los meses de febrero o marzo y, por tanto, la mayor parte de la variabilidad de la serie se debe indudablemente a este efecto estacional. Es posible que exista autocorrelación ya que los valores de cada mes parecen depender del anterior o de los meses anteriores. Los picos van cambiando en el tiempo, lo que podría indicar una serie no estacionaria. No parece haber una tendencia clara, oscilando en torno a un valor medio razonablemente estable, por lo que es posible considerar –con dudas ya que la serie es muy corta– que se trata de una serie estacionaria. Si observásemos una tendencia creciente o decreciente, sería necesario diferenciar la serie.

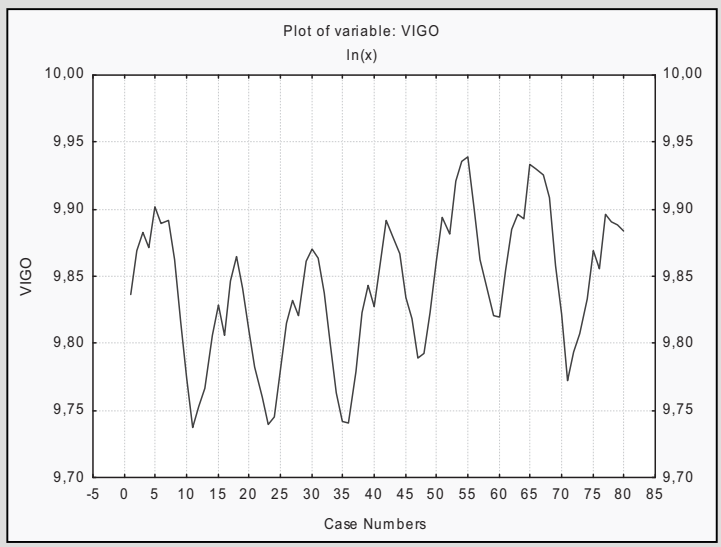


Paso 5. La amplitud de los picos estacionales parece que es igual, pero de no ser así sería necesario realizar una transformación. Para ello en la pestaña «x=f(x)» se selecciona alguna función de transformación adecuada, como por ejemplo «Natural log (x=ln(x))» y se pulsa «OK (Transform selected series)».

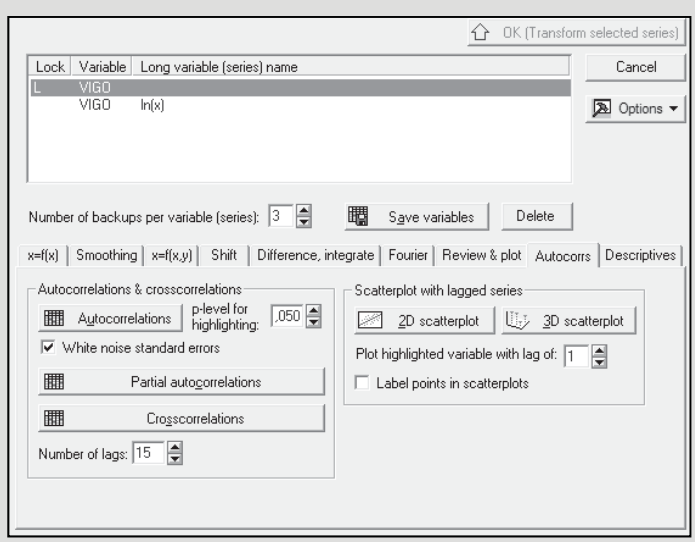


Cuadro XII.1. *(Continuación)* *STATISTICA*

En el gráfico vemos el resultado de la transformación. Si la serie original no tiene amplitud constante, debemos encontrar una transformación que tenga esta propiedad, lo que no es necesario en este caso. Por tanto, no es necesario trabajar con la variable transformada logarítmicamente.

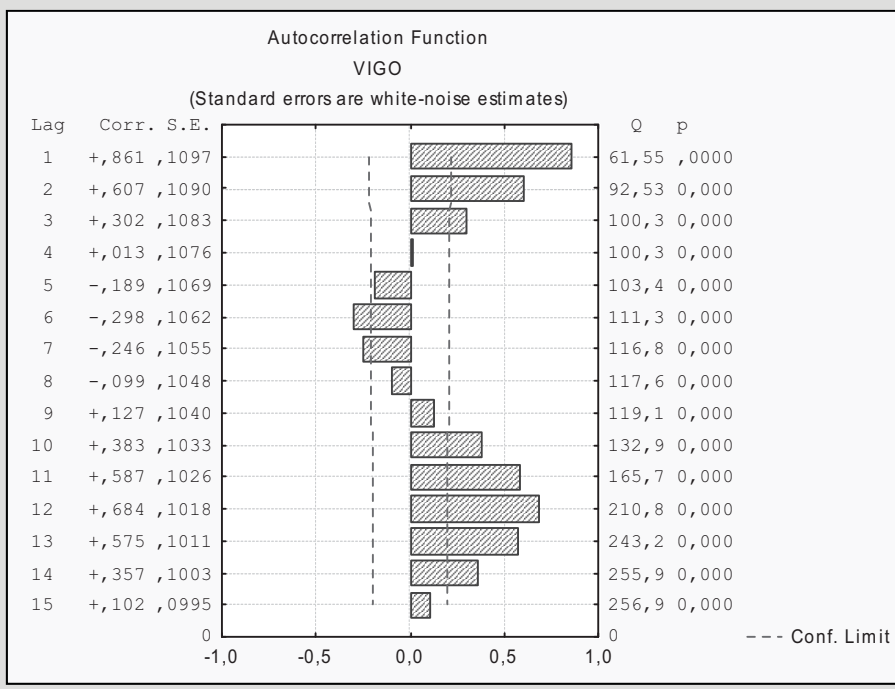


Paso 6. Autocorrelación. Para decidir los valores adecuados de *p* y *q*, es útil estudiar la correlación o asociación entre los valores de la serie en cada período y los valores en los períodos anteriores, es decir la relación entre la serie y ella misma (autocorrelación) desplazada uno o más períodos. Para ver la autocorrelación nos aseguramos de que esté seleccionada la variable «Vigo» y en la pestaña «Autocorrs» pulsamos en «Autocorrelations».



Cuadro XII.1. *(Continuación)*[STATISTICA]

Paso 7. La función de autocorrelación simple (*ACF*), que se observa en el siguiente gráfico, muestra el conjunto de coeficientes de autocorrelación, e indica hasta que punto la variable depende de ella misma. Cuando el valor de la serie Y_t depende fuertemente del valor anterior Y_{t-1} existirá un coeficiente de correlación alto R_1 entre la serie y la misma serie desplazada un período, pero dado que Y_{t-1} depende a su vez del valor anterior a él Y_{t-2}, también será alto el coeficiente de correlación con la serie desplazada dos períodos, R_2, y así sucesivamente.

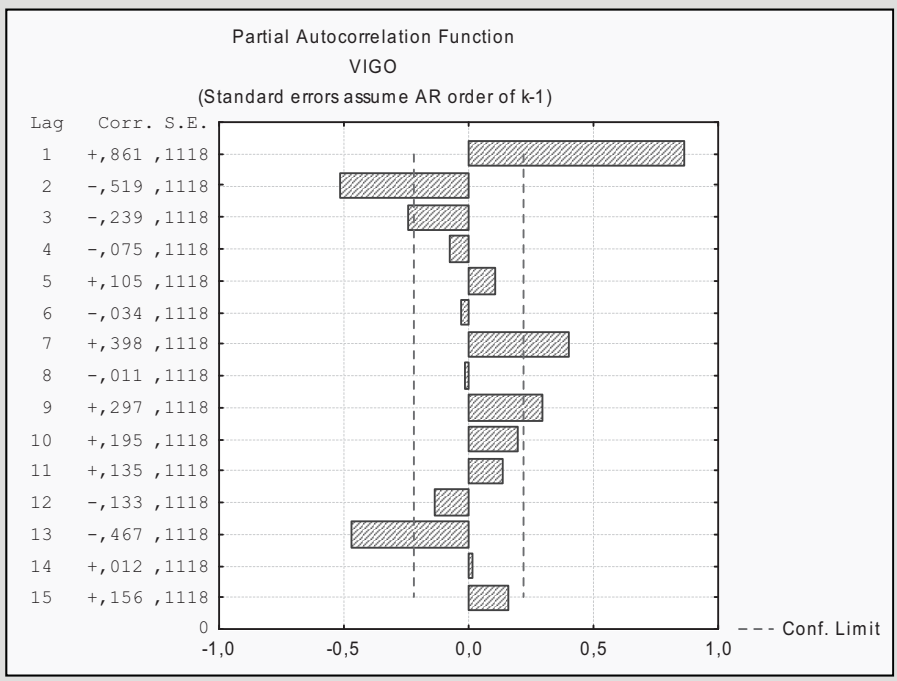


Si en la *AFC* los coeficientes (positivos o negativos) disminuyen gradualmente, como es el caso que muestra la figura, esto indica que la serie depende fuertemente de los valores de la misma serie próximos en el tiempo, pero esta dependencia se va atenuando a medida que nos alejamos. Se observa como las correlaciones más altas son para una diferencia de 1 mes (r = 0,861) y de 12 meses (r = 0,684).

Sin embargo, es posible que una parte de esa dependencia que mide el coeficiente de correlación simple sea en realidad indirecta, a través de los valores intermedios, por lo que solamente los valores más próximos al actual deben ser considerados (tal vez uno solo). Por ello, el coeficiente de correlación parcial (*ACF* parcial o *ACFP*) es más adecuado para analizar estas relaciones de dependencia que el coeficiente de correlación simple, ya que al eliminar la influencia de las restantes variables (los valores intermedios) aísla el efecto propio de cada retardo sobre el valor actual de la serie, y permite estudiar la correlación entre la serie y la misma serie retardada sin incluir los efectos indirectos.

Cuadro XII.1. *(Continuación)* STATISTICA

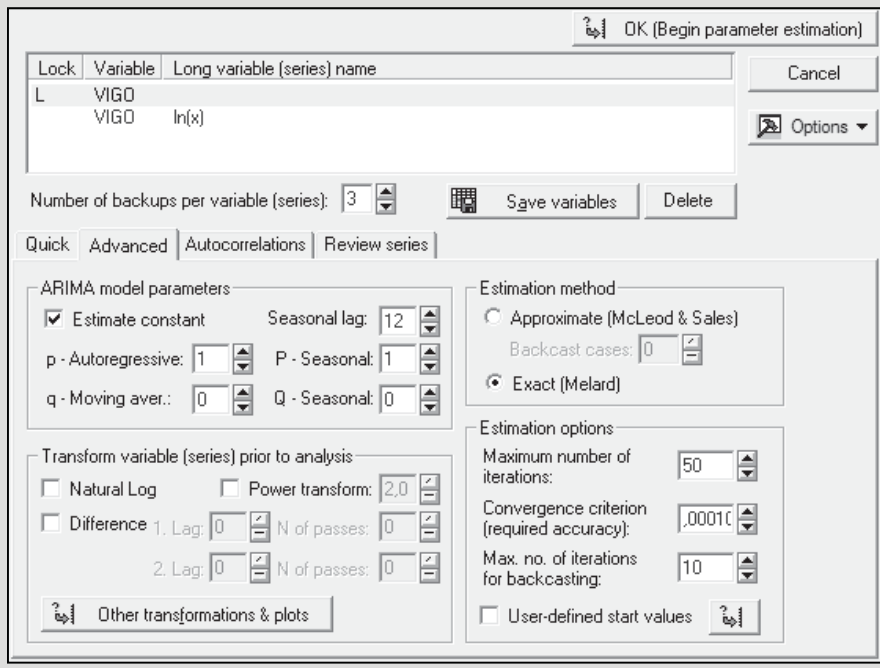
Paso 8. Para calcular *ACF* parcial en el paso 6 se pulsa «Partial autocorrelations». En la siguiente figura se muestra la *ACF* parcial correspondiente a los mismos datos que la anterior, en la que se puede apreciar como el primer coeficiente es mucho más importante que los demás, lo que sugiere un modelo con $p = 1$.



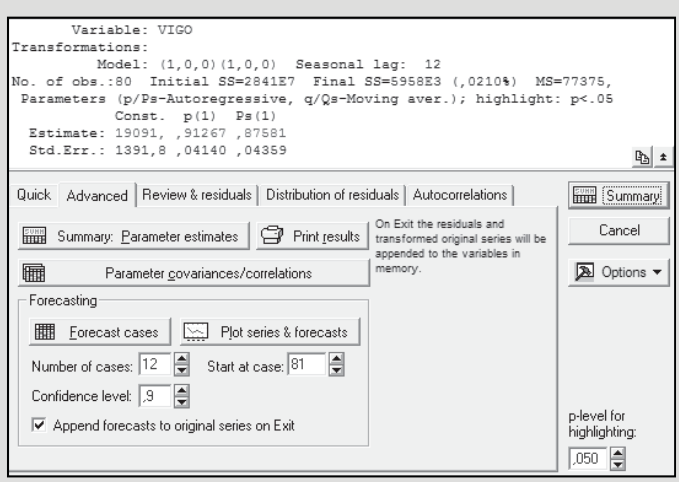
Paso 9. Estacionalidad y desestacionalización. Tanto el gráfico secuencial de la serie (paso 4) como la función de autocorrelación (paso 7) permiten apreciar claramente la estacionalidad de la serie (se repite la misma pauta todos los años). Se observa en *ACF*, aunque no tan evidente en *ACF* parcial, que existe una clara estacionalidad cada 12 meses.

Paso 10. En la ventana del paso 6 se pulsa en «Cancel» y aparece la ventana del paso 2, en la cual entramos en «ARIMA & autocorrelation functions», donde nos aparece la siguiente ventana. Aplicaremos el modelo ARIMA directamente a la variable no transformada. En la pestaña «Advanced» introducimos los parámetros del modelo ARIMA, buscando inicialmente el modelo más sencillo posible. De acuerdo con las funciones de autocorrelación parece adecuado establecer $p = 1$; además la serie es estacionaria, por lo que haremos $d = 0$ (ninguna transformación). En cuanto a q, empezaremos con el valor más simple, $q = 0$. El modelo será por lo tanto ARIMA (1,0,0). En lo que se refiere a la parte estacional del modelo ARIMA, puesto que sabemos por el gráfico de secuencia que existe una clara estacionalidad anual, utilizaremos en esa parte el mismo modelo simple (1,0,0). Marcaremos además «Estimate constant» y en el método de estimación «Exact (Melard)». Pulsamos «OK (begin parameter estimation)».

Cuadro XII.1. *(Continuación)* STATISTICA

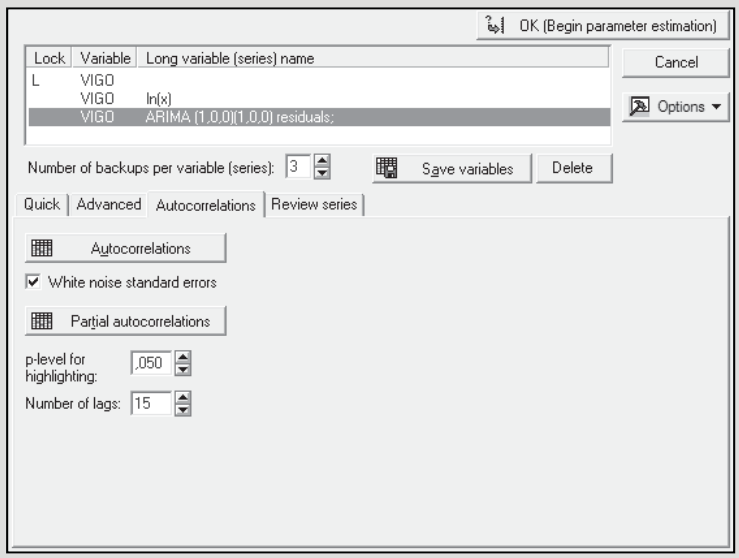


Paso 11. Comprobamos que los parámetros estimados aparecen resaltados (en color rojo), indicando que son estadísticamente significativos. Por otra parte, la suma de cuadrados finales o residuales es solamente el 0,021% de la suma de cuadrados iniciales, lo que muestra que el modelo se ajusta bien a los datos y los residuos son pequeños. Si alguno de los coeficientes no es significativo, podemos simplificar el modelo eliminando esa componente, o mantenerlo como está; por ejemplo, si sabemos que existe una estacionalidad muy clara, podemos mantener $p = 1$ en la parte estacional aunque no sea significativo estadísticamente.

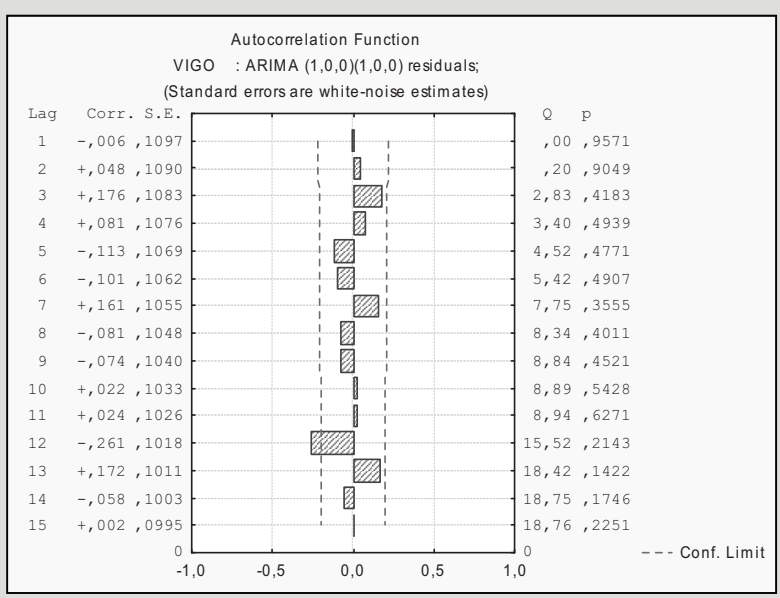


Cuadro XII.1. *(Continuación)* *STATISTICA*

Paso 12. Para averiguar si necesitamos un modelo más complejo construimos la función de autocorrelación de los residuos. Para ello, pulsando en «Cancel» volvemos a la siguiente ventana, en la que seleccionamos la variable «VIGO ARIMA(1,0,0)(1,0,0) Residuals», que contiene los residuos. Pulsamos sobre la pestaña «Autocorrelations», y dentro de ella de nuevo en «Autocorrelations» para obtener la función de autocorrelación simple.



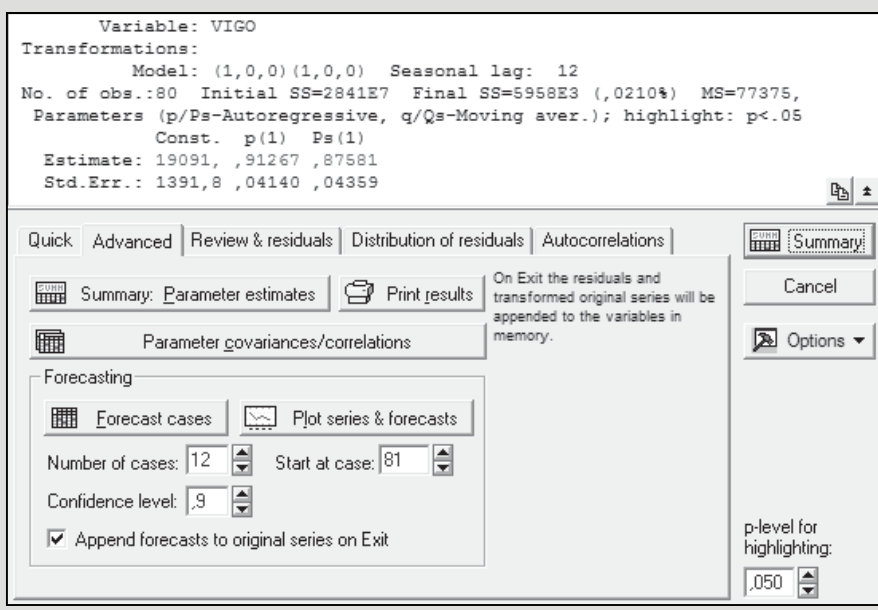
Paso 13. Aparecerá la función de autocorrelación simple de los residuos, que se muestra a continuación.



Cuadro XII.1. *(Continuación)*[STATISTICA]

Como se puede comprobar, no existen autocorrelaciones significativas entre los residuos salvo para el retardo de 12 meses, aunque también este coeficiente de correlación es pequeño (-0,261), por lo que los residuos no contienen una estructura que deba ser explicada mediante un modelo más complicado. En la parte derecha del gráfico anterior aparece –para cada coeficiente de autocorrelación– el estadístico Q de Ljung-Box, junto con su nivel de significación. Un valor *p* menor que 0,05 indicaría una posible estructura de autocorrelaciones, lo que no se observa en este ejemplo. Nuestro modelo ARIMA es adecuado y finaliza la fase de especificación. Ahora podemos aplicar nuestro modelo al análisis de los valores de la serie en el pasado y a la predicción de valores futuros.

Paso 14. Para predecir los valores futuros de la serie volvemos a la pestaña «Advanced» en la que pulsamos en «Forecast cases», después de establecer el número de valores que queremos predecir. Por defecto estima un período completo de estacionalidad (en este caso 12 meses) a partir del último dato conocido. En general los modelos ARIMA solo permiten estimaciones razonablemente fiables a corto plazo.

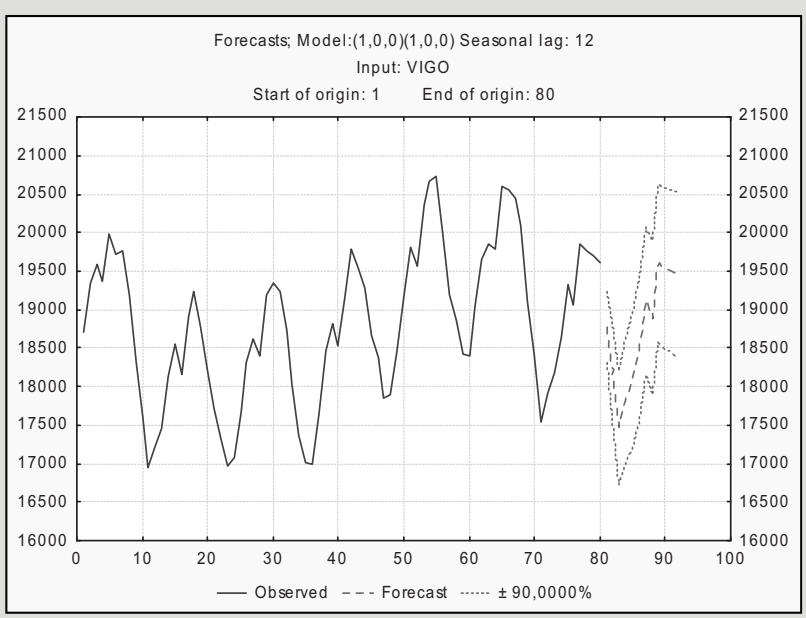


Paso 15. Aparecerán los valores pronosticados por el modelo, junto con un intervalo de estimación que mide de algún modo la fiabilidad o el margen de error que podemos atribuir a estas predicciones. A menudo los pronósticos son el resultado de mayor interés en la aplicación de los modelos ARIMA. El número de caso «CaseNo.», a la izquierda de los valores de predicción «Forecast» nos permite identificar el período al que corresponde cada valor.

Cuadro XII.1. *(Continuación)* STATISTICA

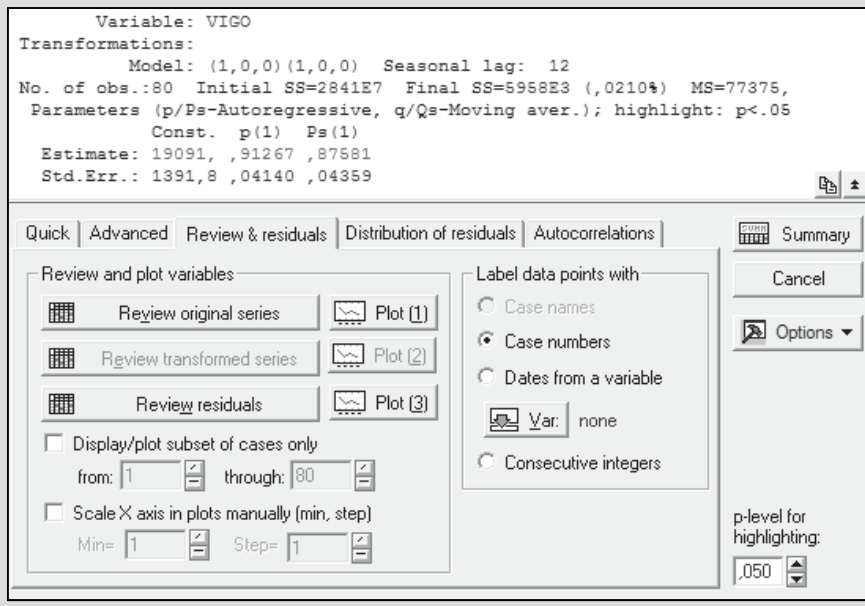
CaseNo.	Forecasts; Model:(1,0,0)(1,0,0) Seasonal lag: 12 (Cuadro XII.1.sta) Input: VIGO Start of origin: 1 End of origin: 80 Forecast	Lower 90,0000%	Upper 90,0000%	Std.Err.
81	18782,43	18319,32	19245,54	278,1633
82	18213,30	17586,30	18840,29	376,5972
83	17469,16	16733,01	18205,32	442,1658
84	17818,43	17002,41	18634,44	490,1312
85	18061,42	17184,42	18938,41	526,7609
86	18492,59	17567,86	19417,32	555,4312
87	19116,49	18153,81	20079,17	578,2284
88	18893,67	17900,48	19886,86	596,5530
89	19603,05	18585,14	20620,95	611,3977
90	19530,78	18492,73	20568,82	623,4930
91	19497,61	18443,09	20552,13	633,3917
92	19432,71	18364,65	20500,77	641,5204

Paso 16. Podemos observar los valores que predice el modelo, y representarlos gráficamente, pulsando en «Plot series & forecasts» en la ventana del paso 14. En el gráfico vemos como la serie de valores pronosticados (en color rojo) sigue la evolución aproximada de los datos conocidos, reproduciendo el esquema estacional. Obsérvese por ejemplo el descenso del paro en el mes de diciembre (atribuible a la campaña comercial de Navidad), que se repite todos los años rompiendo la racha ascendente de septiembre a marzo, y que aparece claramente también en la predicción. En color verde (línea punteada) se muestran los límites de confianza, que progresivamente se van separando de la predicción indicando así la menor precisión que se obtiene a medida que nos alejamos en el tiempo.



Cuadro XII.1. *(Continuación)*[STATISTICA]

Paso 17. Los residuos pueden ser utilizados para estudiar los valores pasados de la serie. Se selecciona la pestaña «Review & residuals» y pulsamos en «Review residuals».



Paso 18. En estos datos de desempleo un residuo grande puede ser debido al cierre de una empresa, o al efecto de un plan puntual de contratación pública en la lucha contra el paro.

	Listing of variable: VIGO (Cuadro XII.1.sta) ARIMA (1,0,0)(1,0,0) residuals;
Case	Value
1	-55,825
2	297,433
3	150,193
4	-67,198
5	329,123
6	-53,721
7	88,880
8	-213,693
9	-395,278
10	-432,525
11	-476,231
12	-145,802
13	-224,106
14	22,980
15	90,745

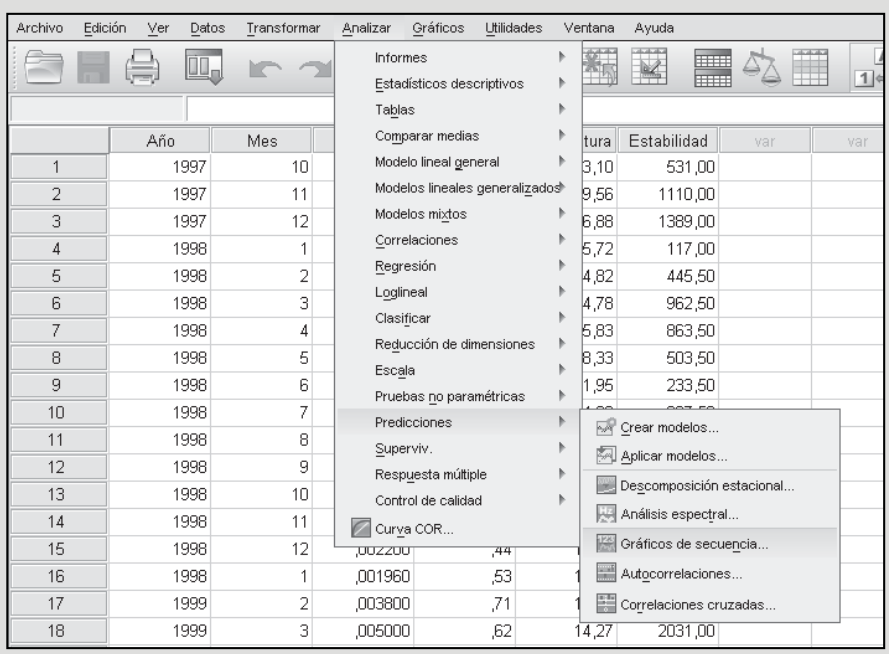
Cuadro XII.1. *(Continuación)*[SPSS]

Aplicación con SPSS.

EJEMPLO. Se va a utilizar el mismo ejemplo que en la aplicación con R. La serie temporal que se va a analizar está en el archivo **Cuadro XII.1.sav**. Los datos muestran el factor de condición mensual (*FC*) de la anchoa (*Engraulis encrasicolus)* a lo largo de varios años. *FC* es un indicador del estado reproductivo de los individuos.

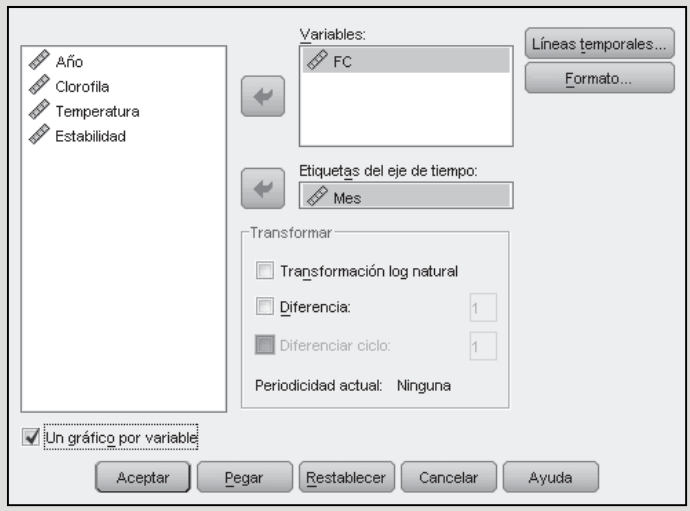
El objetivo es determinar el efecto de los valores medios mensuales de la temperatura (en ºC), clorofila (en mg m^{-3}) y estabilidad de la columna de agua (en m^3s^{-3}) sobre el estado reproductivo de esta especie. Como *FC* y el resto de variables pueden covariar con el tiempo, es necesario cuantificar el posible efecto del tiempo sobre *FC*, en otras palabras, introducir el potencial efecto del tiempo sobre *FC* en el modelo, para poder determinar el efecto real de las variables mencionadas anteriormente sobre *FC*.

Paso 1. En primer lugar representamos la variable *FC* para tener una idea de como varía en el tiempo. Para ello seleccionamos «Analizar» ⇨ «Predicciones» ⇨ «Gráficos de secuencia».

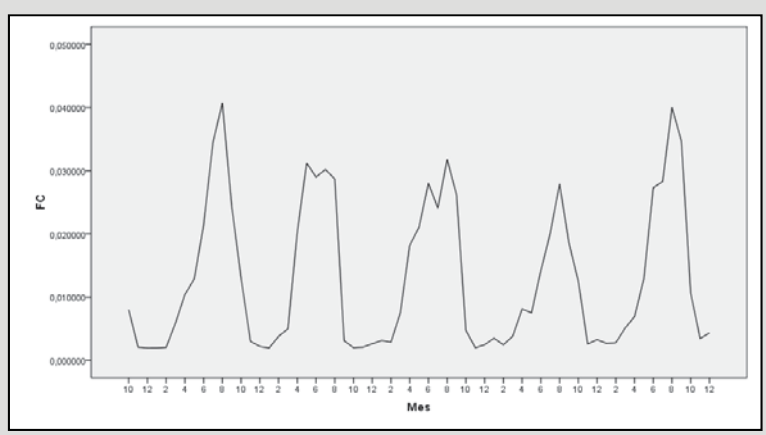


Paso 2. Aparece la siguiente ventana donde se especifica la variable a representar, en este caso *FC*, frente al tiempo expresado en meses.

Cuadro XII.1. *(Continuación)* *SPSS*

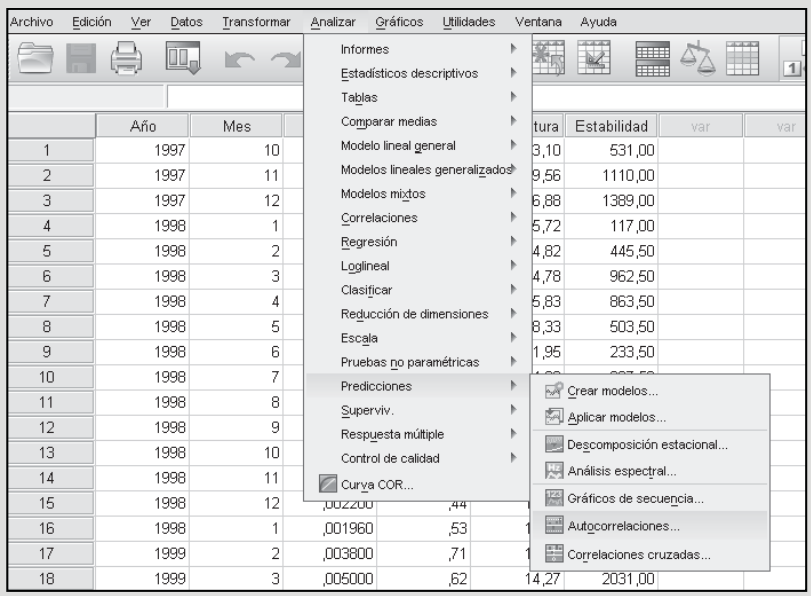


Paso 3. En la gráfica que se obtiene se pueden observar algunos aspectos importantes para la construcción del modelo ARIMA. Es posible que exista autocorrelación ya que los valores de un mes parecen depender del anterior o anteriores meses. La serie parece estacionaria, en el sentido de que no se observa ninguna tendencia creciente ni decreciente en el tiempo. Por último, se ve el carácter estacional de la serie, que repite el mismo patrón cada año con valores altos en la época de verano y bajos en invierno. La amplitud estacional parece ser muy parecida. Si las amplitudes estacionales variasen en el tiempo, sería conveniente probar a realizar el modelo con los datos transformados.

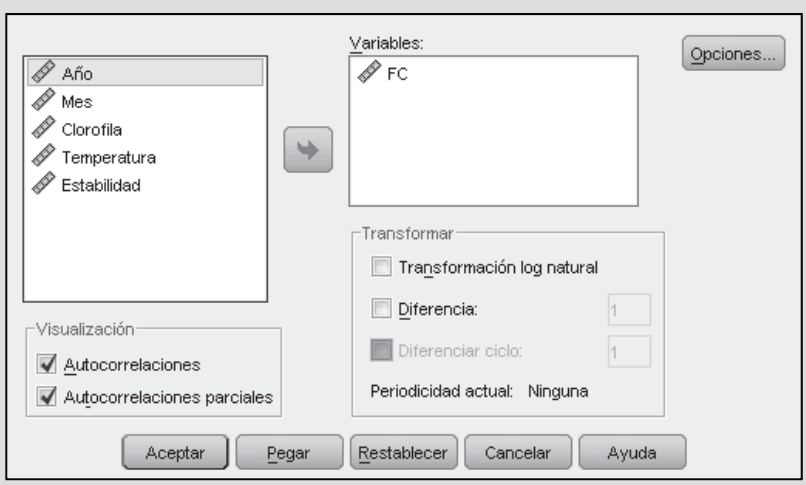


Paso 4. Autocorrelación. Para decidir los valores adecuados de *p* y *q*, es útil estudiar la correlación o asociación entre los valores de la serie en cada período y los valores en los períodos anteriores, es decir la relación entre la serie y ella misma (autocorrelación) desplazada uno o más períodos. Para ver la autocorrelación seleccionamos «Analizar» ⇨ «Predicciones» ⇨ «Autocorrelaciones...».

Cuadro XII.1. *(Continuación)* SPSS

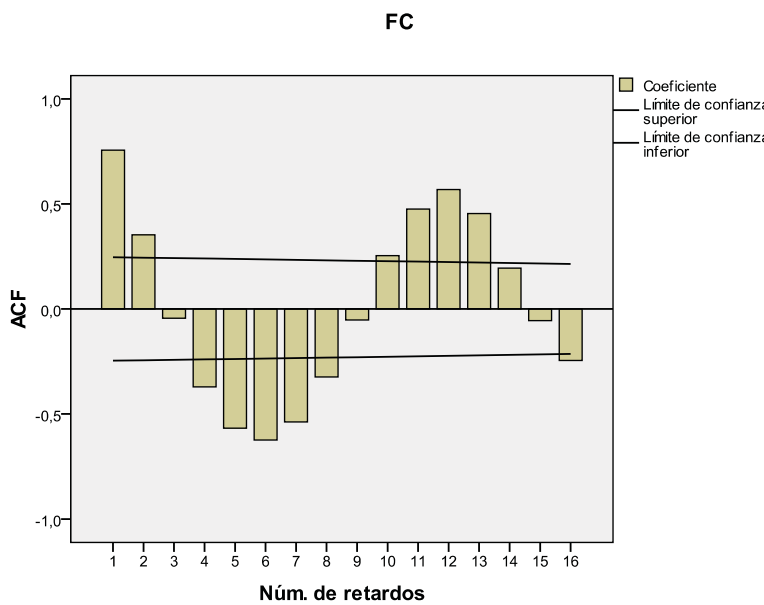


Paso 5. Aparece la siguiente ventana donde seleccionamos *FC* y nos aseguramos de que en «Visualización» estén marcadas las opciones «Autocorrelaciones» y «Autocorrelaciones parciales».



La función de autocorrelación simple (*ACF*), que se observa en el siguiente gráfico y tabla, muestra el conjunto de coeficientes de autocorrelación, e indica hasta que punto la variable depende de ella misma. Cuando el valor de la serie Y_t depende fuertemente del valor anterior Y_{t-1} existirá un coeficiente de correlación alto R_1 entre la serie y la misma serie desplazada un período, pero dado que Y_{t-1} depende a su vez del valor anterior a él Y_{t-2}, también será alto el coeficiente de correlación con la serie desplazada dos períodos, R_2, y así sucesivamente.

Cuadro XII.1. *(Continuación)* SPSS



Autocorrelaciones

Serie:FC

Retardo	Autocorrelación	Típ. Error[a]	Estadístico de Box-Ljung		
			Valor	gl	Sig.[b]
1	,756	,123	37,773	1	,000
2	,353	,122	46,132	2	,000
3	-,044	,121	46,266	3	,000
4	-,371	,120	55,821	4	,000
5	-,568	,119	78,579	5	,000
6	-,623	,118	106,507	6	,000
7	-,537	,117	127,628	7	,000
8	-,324	,116	135,440	8	,000
9	-,052	,115	135,646	9	,000
10	,254	,114	140,611	10	,000
11	,476	,113	158,424	11	,000
12	,569	,112	184,382	12	,000
13	,454	,110	201,297	13	,000
14	,194	,109	204,452	14	,000
15	-,056	,108	204,716	15	,000
16	-,245	,107	209,943	16	,000

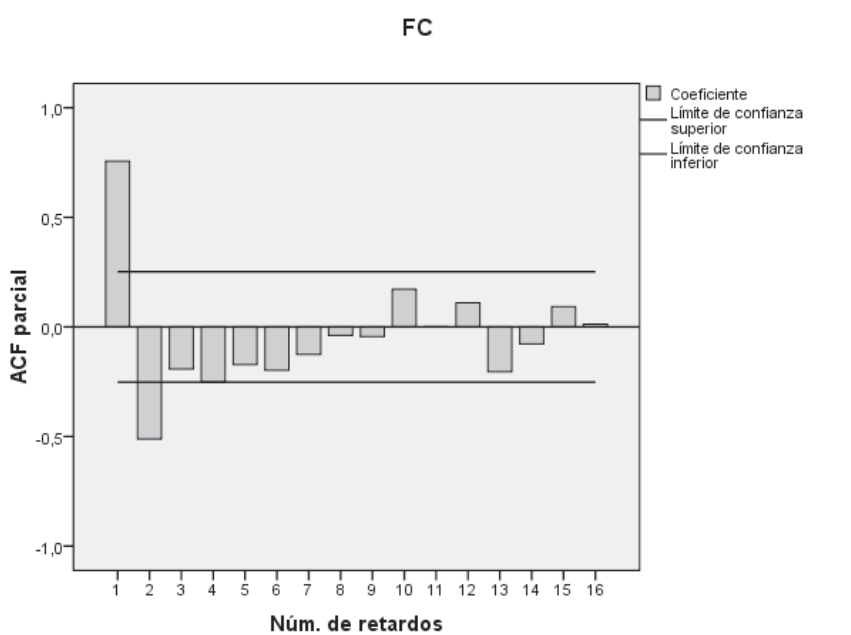
a. El proceso subyacente asumido es la independencia (ruido blanco).

Cuadro XII.1. *(Continuación)* *SPSS*

Si en la *AFC* los coeficientes (positivos o negativos) disminuyen gradualmente, cómo es el caso que muestra la figura y tabla anteriores, esto indica que la serie depende fuertemente de los valores de la misma serie próximos en el tiempo, pero esta dependencia se va atenuando a medida que nos alejamos.

Sin embargo, es posible que una parte de esa dependencia que mide el coeficiente de correlación simple sea en realidad indirecta, a través de los valores intermedios, por lo que solamente los valores más próximos al actual deben ser considerados (tal vez uno solo). Por ello, el coeficiente de correlación parcial (*ACF* parcial o *ACFP*) es más adecuado para analizar estas relaciones de dependencia que el coeficiente de correlación simple, ya que al eliminar la influencia de las restantes variables (los valores intermedios) aísla el efecto propio de cada retardo sobre el valor actual de la serie, y permite estudiar la correlación entre la serie y la misma serie retardada sin incluir los efectos indirectos.

En las siguientes figura y tabla se muestra la *ACF* parcial correspondiente a los mismos datos que la anterior, en las que se puede apreciar como en realidad solamente el primero o los dos primeros coeficientes de autocorrelación (primer o segundo retardo) son significativos.



Cuadro XII.1. *(Continuación)* *SPSS*

Autocorrelaciones parciales

Serie:FC

Retardo	Autocorrelación parcial	Típ. Error
1	,756	,126
2	-,512	,126
3	-,192	,126
4	-,251	,126
5	-,172	,126
6	-,198	,126
7	-,125	,126
8	-,038	,126
9	-,044	,126
10	,172	,126
11	,000	,126
12	,110	,126
13	-,204	,126
14	-,078	,126
15	,092	,126
16	,012	,126

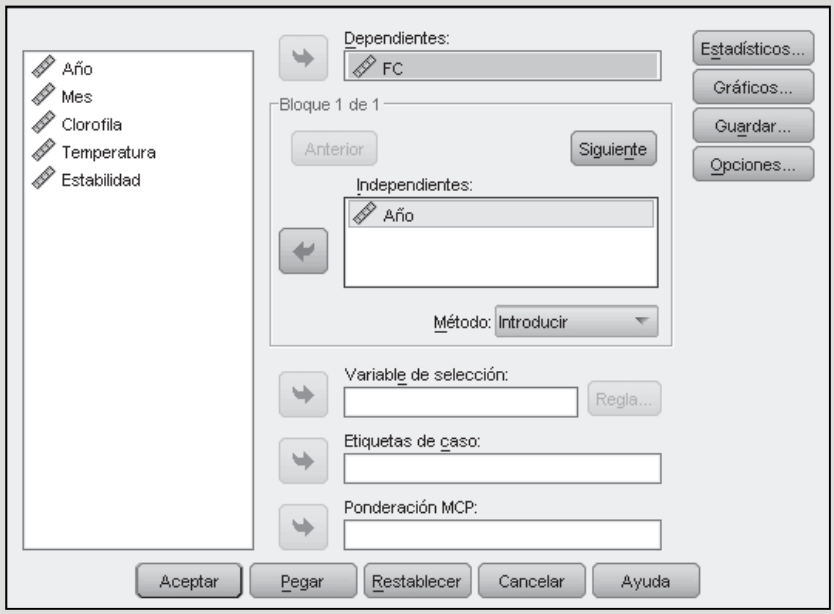
La *ACFP* nos puede ayudar a elegir los valores adecuados de *p*, aunque no existen normas fijas. Si la *ACF* decae de forma sostenida y la *ACFP* tiene los *k* primeros valores significativos, puede indicar un modelo con $p = k$, siendo *k* el número de retardos que se considere. Como sólo el primero o los dos primeros coeficientes de autocorrelación son importantes (superiores a 0,5 o próximos a este valor), *p* debería ser igual a 1 ó 2 en nuestro modelo ARIMA.

El valor de la *ACF* sirve para tener una aproximación del valor *q*, aunque tampoco existen unas normas fijas. Si la *ACFP* decae exponencialmente y *ACF* tiene los *k* primeros valores significativos, ello puede indicar un modelo con $q = k$. En nuestro caso como solo el primer retardo es importante en la *ACF*, *q* podría ser igual a 1.

Paso 6. Determinar si la serie es estacionaria. El valor de *d* debe ser cero si la serie es estacionaria, y si no lo es tomará un valor mayor que cero (1 ó 2, en general).

Para determinar si la serie es estacionaria o, por el contrario, crece o decrece con el tiempo, el método más común es ver si la pendiente de la serie con el tiempo es significativa. Para ello, como se explicó en el capítulo de regresiones en «Analizar» ⇨ «Regresión» ⇨ «Lineales...», aparece la siguiente ventana donde se selecciona *FC* como variable dependiente y «Año» como independiente.

Cuadro XII.1. *(Continuación)* SPSS



El resultado que se obtiene muestra que la relación entre ambas variables no es significativa con una p = 0,629. Por tanto, la serie es estacionaria, ni crece ni disminuye con el tiempo y d debería ser igual a 0.

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	,000	1	,000	,236	,629[a]
	Residual	,009	61	,000		
	Total	,009	62			

a. Variables predictoras: (Constante), Año

b. Variable dependiente: FC

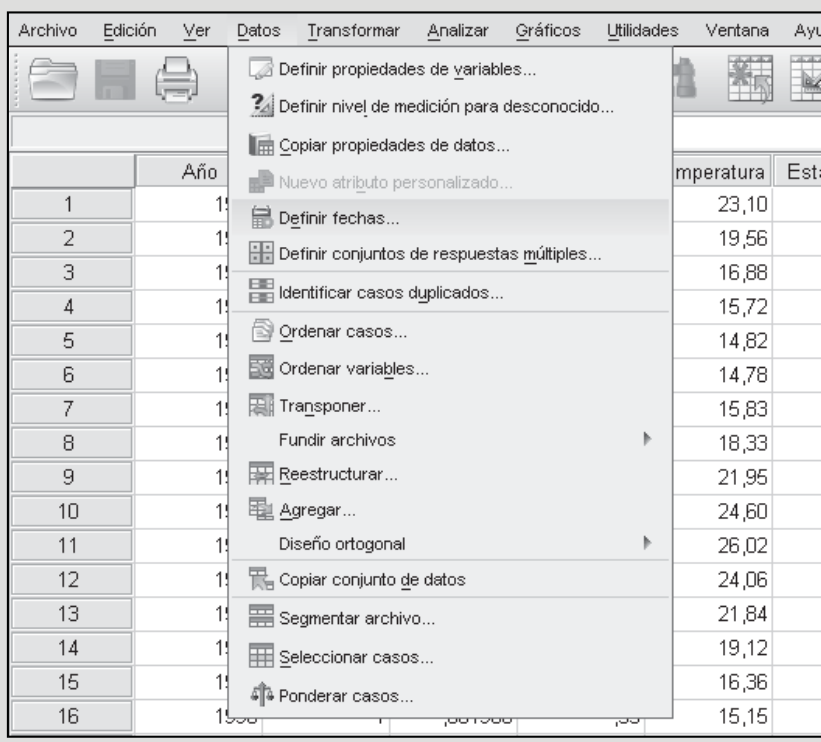
Paso 7. Estacionalidad. Tanto el gráfico secuencial de la serie (Paso 3) como la función de autocorrelación (Paso 5) permiten apreciar claramente la estacionalidad de la serie (se repite la misma pauta todos los años). Se observa en *ACF* (Paso 5) aunque no tan evidente en *ACF* parcial, el efecto estacional, ya que los coeficientes aumentan en las proximidades del retardo 6 y 12, lo cual indica que el valor de la variable está relacionado positivamente con el valor que tenía en el mismo mes el año anterior y 6 meses antes, pero en este segundo caso con una relación inversa.

Por lo tanto, la estacionalidad debe ser tenida en cuenta en la construcción del modelo.

Cuadro XII.1. *(Continuación)* [SPSS]

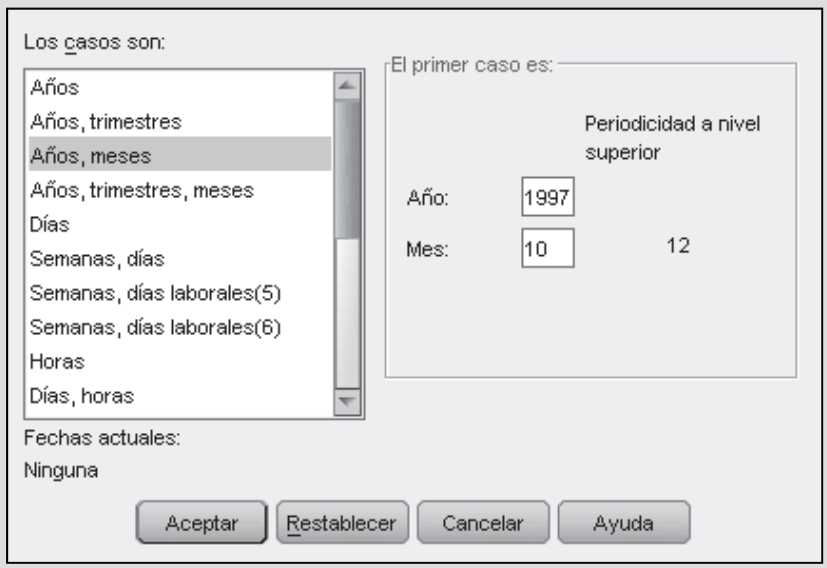
Paso 8. Creación del modelo. El procedimiento tradicional consiste en construir diferentes modelos de prueba utilizando distintos valores de *p*, *d* y *q*, tanto para la parte regular del modelo como para la parte estacional, y elegir el más adecuado utilizando para ello los diferentes criterios de diagnóstico, o medidas de bondad de ajuste disponibles. La experiencia permite reducir razonablemente el campo de búsqueda, especialmente si aplicamos estos modelos de forma reiterada a datos similares o conocidos, en los cuales ya sabemos de forma aproximada qué modelo funciona mejor. Las indicaciones mostradas anteriormente permiten acotar la búsqueda de los valores. Los valores de *p*, *d* y *q* suelen ser pequeños (0, 1 ó 2), lo que nos permite establecer un conjunto reducido de modelos de prueba, combinando los diferentes valores de los parámetros.

Paso 9. Aunque ya existan unas variables informando sobre el tiempo, en este ejemplo «Año» y «Mes», hay que definir unas nuevas para poder realizar el modelo. Para ello vamos a «Datos» ⇨ «Definir fechas...».

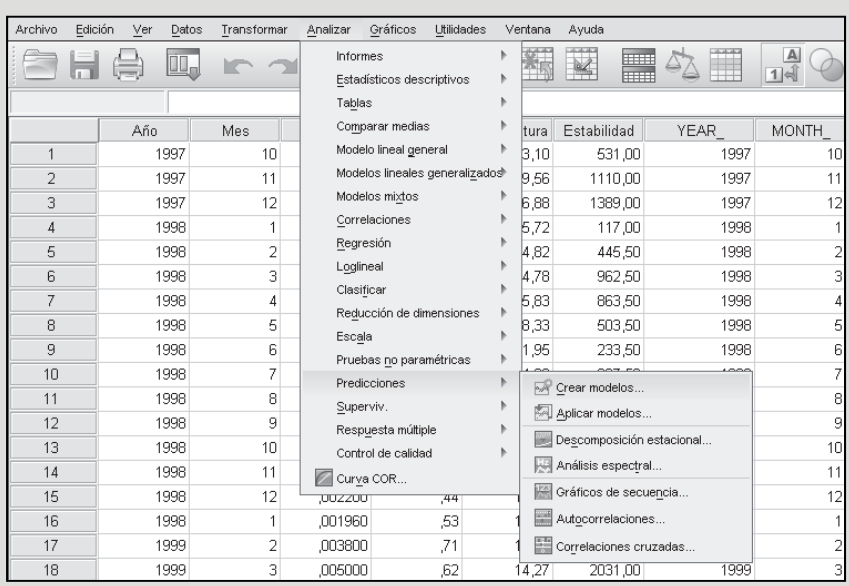


Paso 10. Aparece la siguiente ventana donde elegimos «Años, meses» ya que nuestra serie temporal esta expresada en esas unidades de tiempo y definimos el inicio de la serie, en este caso el año 1997 y el mes 10. Luego aparecerán nuevas columnas, «YEAR», «MONTH» y «DATE».

Cuadro XII.1. *(Continuación)* SPSS

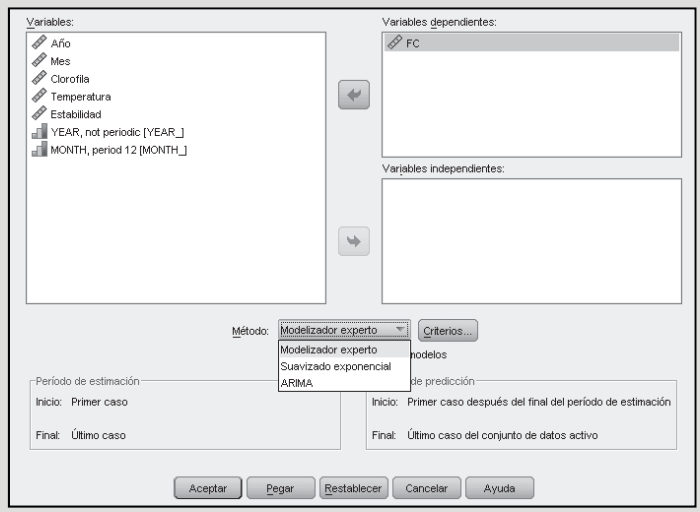


Paso 11. SPSS incluye en sus últimas versiones un modelizador experto, que facilita esta fase de búsqueda de los valores *p*, *d* y *q* más adecuados para el modelo, escogiendo entre todos los modelos razonables aquel que mejor cumple los distintos criterios de ajuste para cada serie. Sin embargo, como veremos más adelante, siempre es conveniente realizar pruebas, con diferentes valores *p*, *d* y *q*, para ver si el modelo escogido por el sistema experto es el más adecuado. Para crear el modelo hay que ir a «Analizar» ⇨ «Predicciones» ⇨ «Crear modelos».

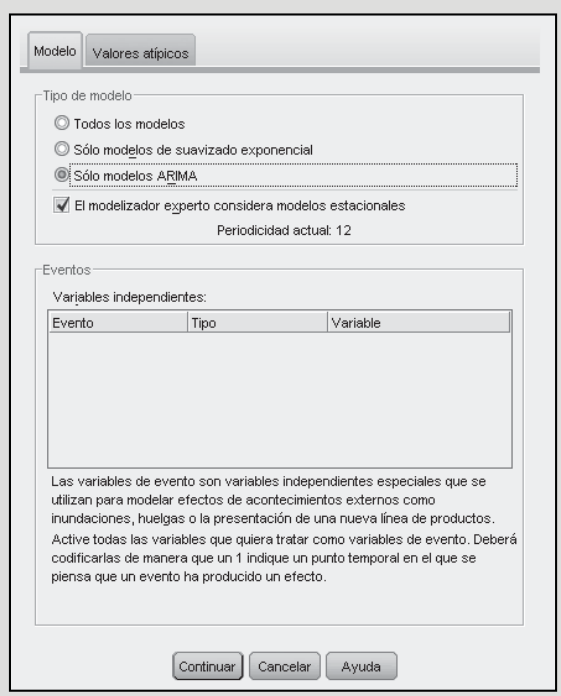


Cuadro XII.1. *(Continuación)* *SPSS*

Paso 12. Aparece la siguiente ventana, donde en la sección «Variables dependientes:» hay que especificar *FC* y en «Método» hay que seleccionar «Modelizador experto».

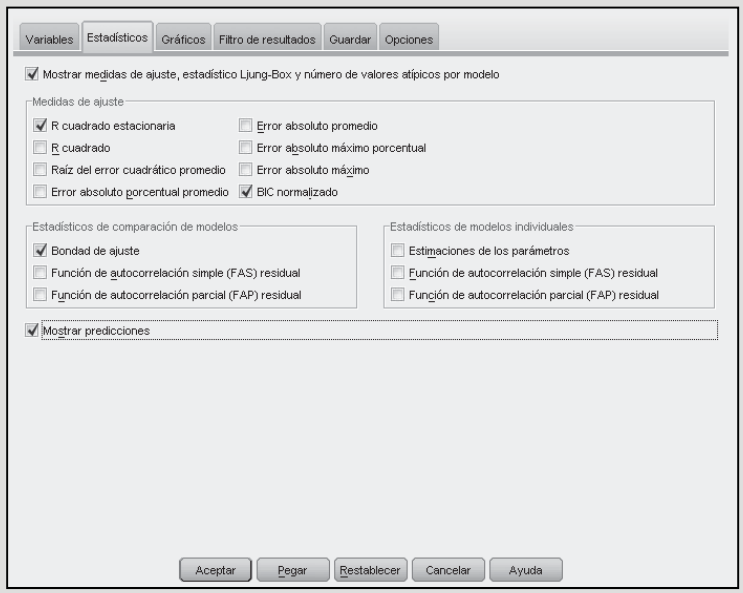


Paso 13. Posteriormente en la ventana anterior hay que entrar en «Criterios» y en la ventana siguiente que aparece seleccionar «Sólo modelos ARIMA».

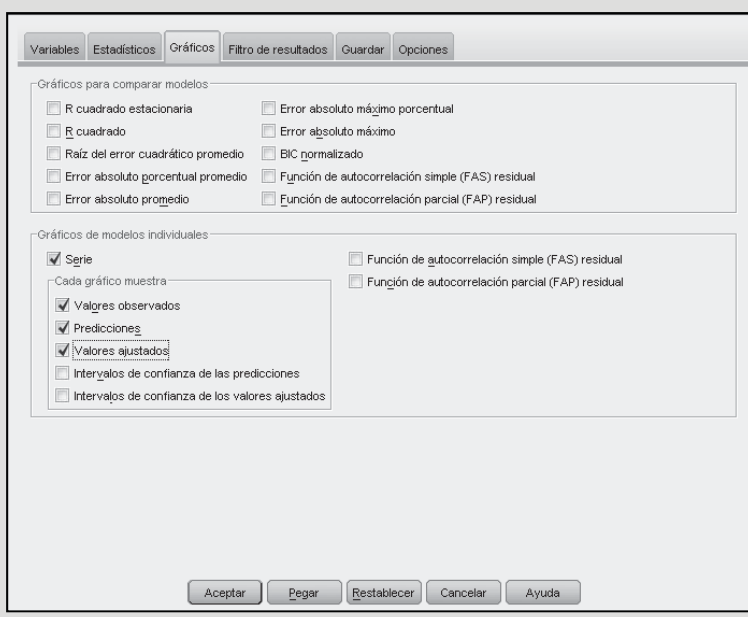

Modelo
Valores atípicos
Tipo de modelo
Todos los modelos
Sólo modelos de suavizado exponencial
Sólo modelos ARIMA
El modelizador experto considera modelos estacionales
Periodicidad actual: 12
Eventos
Variables independientes:
Evento
Tipo
Variable
Las variables de evento son variables independientes especiales que se utilizan para modelar efectos de acontecimientos externos como inundaciones, huelgas o la presentación de una nueva línea de productos.
Active todas las variables que quiera tratar como variables de evento. Deberá codificarlas de manera que un 1 indique un punto temporal en el que se piensa que un evento ha producido un efecto.
Continuar
Cancelar
Ayuda

Cuadro XII.1. *(Continuación)* SPSS

Paso 14. Posteriormente en la sección «Estadísticos» hay que seleccionar «R cuadrado estacionaria», «BIC normalizado», «Bondad de ajuste», «Estimaciones de los parámetros» y «Mostrar predicciones».

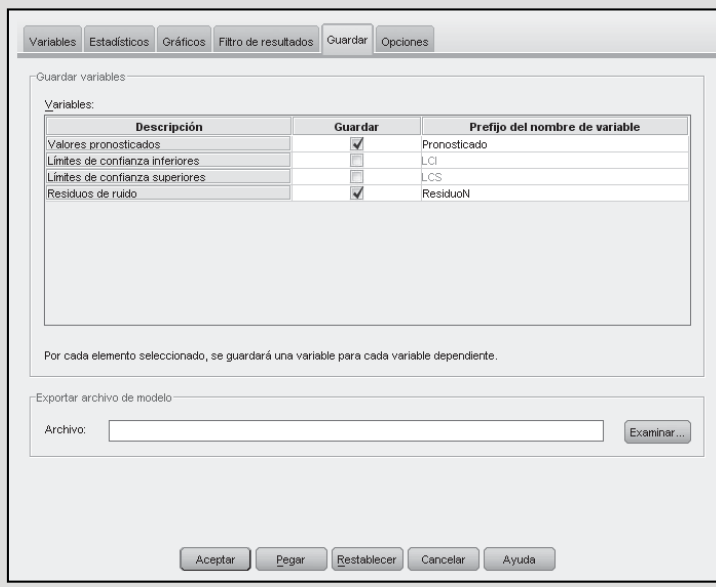


Paso 15. En la sección «Gráficos» se selecciona «Valores observados», «Predicciones» y «Valores ajustados».



Cuadro XII.1. *(Continuación)* *SPSS*

Paso 16. Por último, en la sección «Guardar» se selecciona «Valores pronosticados» y «Residuos de ruido». Los residuos deben ser aleatorios; si no lo son, y conservan alguna estructura, el modelo es incompleto. Una representación gráfica puede ayudar a encontrar la solución, que a menudo consiste en una transformación adecuada de la variable. Por esta razón deben ser guardados para su análisis posterior si es necesario.



Paso 17. Cuando pulsamos «Aceptar» aparecen los resultados del modelo. En primer lugar observamos en «Descripción del modelo» que el sistema experto ha seleccionado para la parte regular los valores (1, 0, 0), es decir, $p = 1$, $d = 0$ y $q = 0$. Esto quiere decir que solo considera significativo un retardo para la componente auto regresiva (p), considera que la serie no crece ni decrece en el tiempo y por ello $d = 0$ y, por último, no considera significativo ningún retardo en la componente de media móvil (q).

Para la componente estacional, el modelizador experto selecciona los valores (0, 1, 0); encuentra por lo tanto que los valores estacionales no son estacionarios y que no existe una estructura estacional autorregresiva o de medias móviles que deba considerarse.

Descripción del modelo

			Tipo de modelo
ID del modelo	FC	Modelo_1	ARIMA(1,0,0)(0,1,0)

Cuadro XII.1. *(Continuación)* SPSS

El estadístico Q de Ljung-Box (o estadístico de Box-Pierce modificado) es adecuado incluso para muestras pequeñas, y se utiliza en un test de falta de ajuste para contrastar la hipótesis nula de que el modelo está correctamente especificado. Un valor *p* menor que 0,05 significa que existe una estructura de correlaciones en los datos que no está adecuadamente recogida en el modelo. Por el contrario, un valor $p > 0,05$ indica que el modelo puede ser correcto. La hipótesis nula que contrasta realmente esta prueba es que en la serie de residuos (diferencias entre los valores de la serie y los valores que predice el modelo) no existe autocorrelación en absoluto. Si existiese autocorrelación en los residuos, ello querría decir que el modelo es incompleto, pues debería incluir un término que explicase esa autocorrelación.

Los resultados de este modelo (1, 0, 0)(0 ,1 ,0) se presentan en la tabla siguiente. El valor $p = 0,089$ del estadístico de Ljung-Box indica que el ajuste puede ser adecuado, aunque el $r^2 = 0,247$, que indica la fracción de variabilidad explicada por el modelo, es bajo.

Estadísticos del modelo

Modelo	Número de predictores	Estadísticos de ajuste del modelo		Ljung-Box Q(18)			Número de valores atípicos
		R-cuadrado estacionaria	BIC normalizado	Estadísticos	GL	Sig.	
FC-Modelo_1	0	,247	-9,669	25,269	17	,089	0

Incrementar los valores de *p*, *q*, y *d* mejora el ajuste del modelo, pero en general es preferible un modelo sencillo, con los valores más pequeños que sea posible de esos parámetros, y esto debe ser tenido en cuenta al elegir el mejor modelo entre varios. Existen algunas medidas de la bondad de ajuste del modelo que tienen en cuenta su complejidad mediante algún tipo de penalización, estableciendo así una protección contra el sobreajuste: se busca el modelo que mejor explica los datos con el mínimo valor posible de los parámetros. La medida más utilizada de este tipo es el "criterio de información de Akaike" (*AIC*) propuesto por Akaike (1974) como una metodología general sencilla para valorar modelos estadísticos. El valor de *AIC* es:

$$AIC = 2k - 2\ln(L)$$

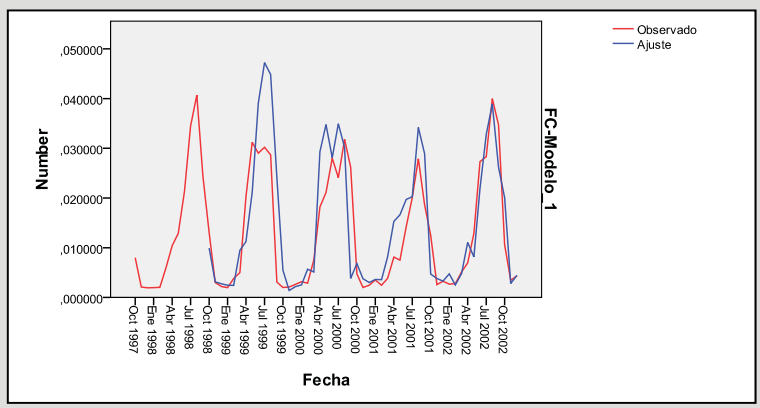
donde *k* es el número de parámetros y *L* la verosimilitud del modelo. Cuanto mayor es la verosimilitud, y cuanto menor es el número de parámetros, más pequeño será el valor de *AIC*. El mejor modelo es el que tiene un menor valor de *AIC*.

SPSS utiliza para las series temporales una medida similar de ajuste global del modelo basado en la misma idea, denominada Criterio de información Bayesiano normalizado (Normalized Bayesian Information Criterion o *BIC* normalizado), que tiene en cuenta la bondad de ajuste y la complejidad del modelo. *BIC* se basa en el error cuadrático medio, incluyendo una penalización para el número de parámetros y la longitud de la serie.

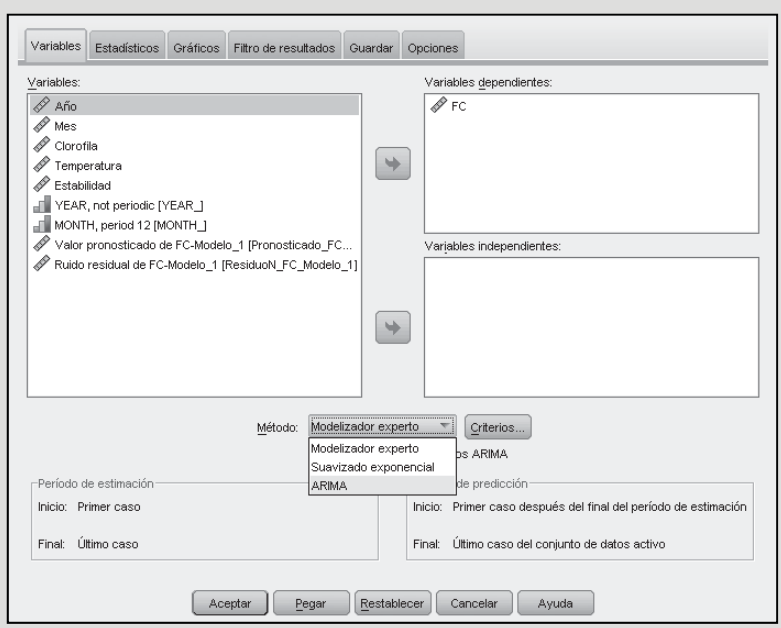
Cuadro XII.1. *(Continuación)* SPSS

El mejor modelo es el que tiene un menor valor *BIC*. No se trata de un contraste, y por lo tanto no existe una hipótesis nula ni un valor *p*, sino solamente el valor del estadístico que únicamente tiene interés para comparar varios modelos entre sí.

Paso 18. En el gráfico, que presenta los valores observados y los estimados por el modelo, se observa como las predicciones del modelo no se ajustan muy bien a los valores reales. Los valores estimados aparecen en una nueva columna en la matriz de datos «Pronosticado_FC Modelo_1».

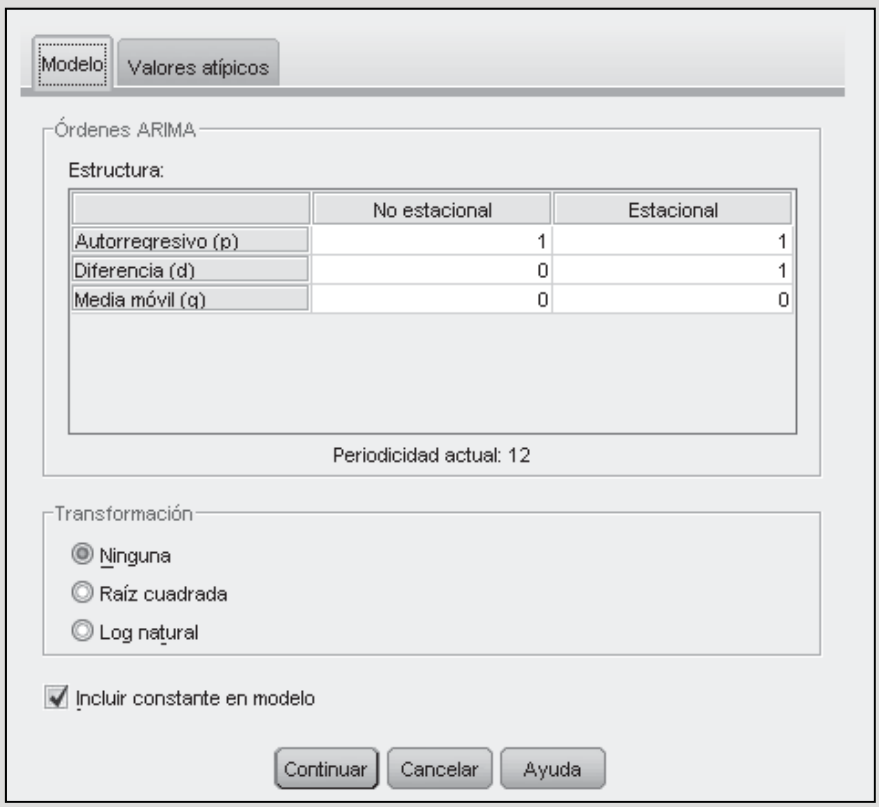


Paso 19. Utilizando como referencia el modelo sugerido por el modelizador experto podemos probar otros similares, introduciendo pequeños cambios. Para ello en la pestaña «Variables», en la sección «Método» hay que seleccionar «ARIMA».



Cuadro XII.1. *(Continuación)* [SPSS]

Paso 20. Como sabemos que existe estacionalidad, podemos intentar un modelo (1, 0, 0)(1, 1, 0), añadiendo un $p = 1$ autorregresivo estacional.



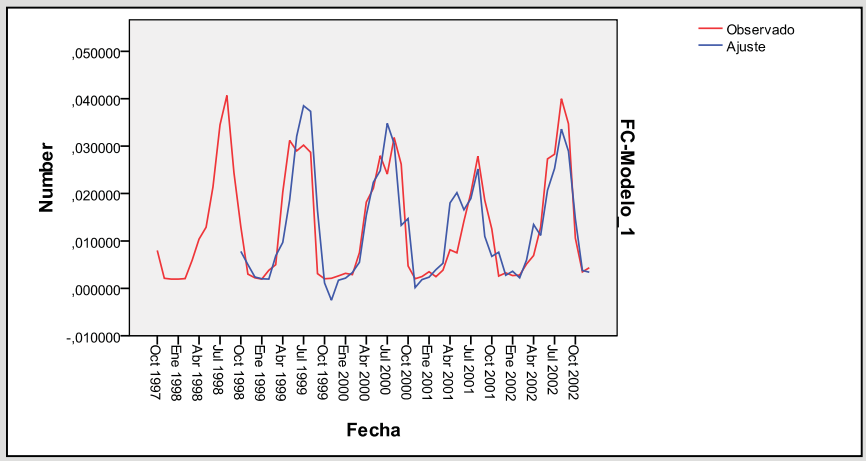
Paso 21. Como podemos observar, el modelo sigue siendo razonablemente completo (P Ljung-Box=0,109), pero hemos mejorado el valor de $r^2 = 0{,}494$, por lo que explica el doble de variabilidad que el anterior y, además, el valor *BIC* es todavía más bajo que en el anterior modelo.

Modelo	Número de predictores	Estadísticos de ajuste del modelo		Ljung-Box Q(18)			Número de valores atípicos
		R-cuadrado estacionaria	BIC normalizado	Estadísticos	GL	Sig.	
FC-Modelo_1	0	,494	-10,039	23,188	16	,109	0

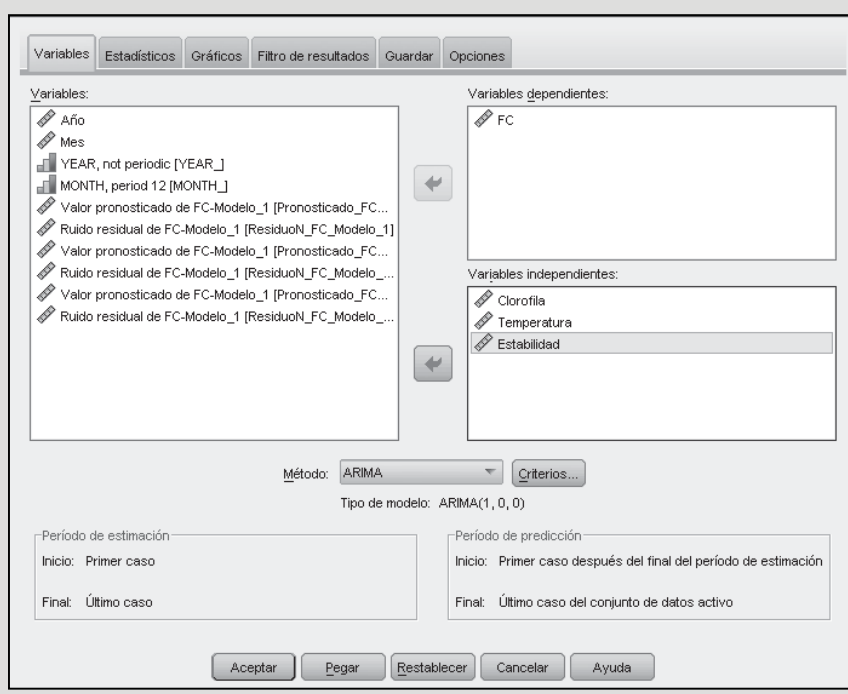
Por lo tanto, el modelizador experto selecciona el modelo más sencillo, que considera razonablemente completo, pero debemos emplear toda la información que conocemos, en este caso la existencia de una clara estacionalidad, para construir un modelo final satisfactorio, utilizando como punto de partida la propuesta del modelizador experto.

Cuadro XII.1. *(Continuación)* [SPSS]

Paso 22. El gráfico muestra que ahora los valores observados y los estimados por el modelo son más parecidos.



Paso 23. Covariables. Una vez que ya se ha obtenido el modelo, en el cual se tiene en cuenta la componente temporal, es posible determinar el efecto de otras variables, lo cual era el objetivo del estudio, sobre la variable dependiente. En la siguiente ventana en la sección «Variables independientes:» hay que incluir «Clorofila», «Temperatura» y «Estabilidad».



Cuadro XII.1. *(Continuación)* SPSS

Paso 24. Los resultados muestran que la predicción del modelo no mejora, ya que el valor de r^2 es muy similar al modelo en el que no se consideraban covariables y, además, el *BIC* no es más pequeño. Por tanto, se concluye que ninguna de las covariables influye en la variable dependiente, es decir, los cambios en el factor de condición mensual (*FC*) de la anchoa (*Engraulis encrasicolus)* no están regulados por cambios en la temperatura, clorofila o en la estabilidad de la columna de agua.

Estadísticos del modelo

Modelo	Número de predictores	Estadísticos de ajuste del modelo		Ljung-Box Q(18)			Número de valores atípicos
		R-cuadrado estacionaria	BIC normalizado	Estadísticos	GL	Sig.	
FC-Modelo_1	3	,496	-9,748	26,148	16	,052	0

En el caso de que existan dudas de que la relación entre las covariables con la variable dependiente sea lineal, es posible transformar las variables y hacer el modelo de nuevo.

Paso 25. Se puede comprobar la Normalidad de los residuos mediante alguna de las técnicas estudiadas en otros capítulos, como por ejemplo en el Cuadro VI.2 en la aplicación con SPSS. Este supuesto se utiliza para encontrar intervalos o límites de confianza para los valores pronosticados, y para los contrastes de hipótesis relativos a los modelos ARIMA. El hecho de que los residuos no tengan una distribución Normal no debe invalidar la aplicación del modelo si la muestra es grande. En el modelo sin tener en cuenta las covariables, los resultados del test de Kolmogorov-Smirnov muestran que los residuos tienen una distribución Normal, ya que la probabilidad es mayor de 0,05 (p = 0,393).

Prueba de Kolmogorov-Smirnov para una muestra

		Ruido residual de FC-Modelo_1
N		51
Parámetros normales[a,,b]	Media	,00001801
	Desviación típica	,005766496
Diferencias más extremas	Absoluta	,126
	Positiva	,097
	Negativa	-,126
Z de Kolmogorov-Smirnov		,900
Sig. asintót. (bilateral)		,393

a. La distribución de contraste es la Normal.

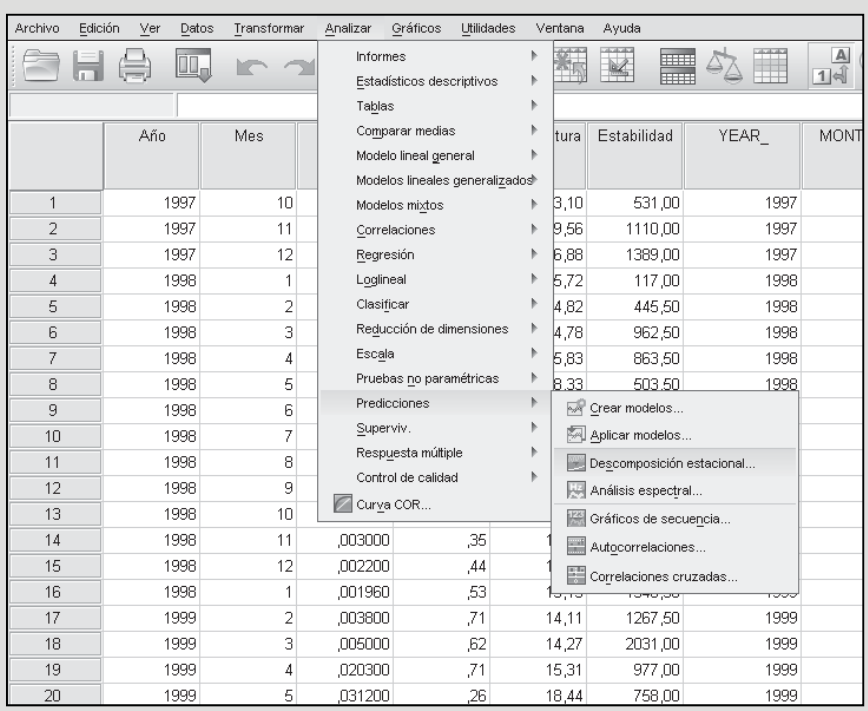
b. Se han calculado a partir de los datos.

Cuadro XII.1. *(Continuación)* [SPSS]

Paso 26. Desestacionalización. En las series que muestran una estacionalidad clara, como en este ejemplo, es decir variaciones regulares de período anual o inferior al año (mensual, semanal, etc.), puede ser difícil interpretar las variaciones a corto plazo. En la figura del paso 3 se observa que en los meses de verano, julio y agosto, se suele dar el máximo de *FC*. La estacionalidad del estado reproductivo es muy acusada y, por lo tanto, la mayor parte de la variabilidad de la serie se debe indudablemente a este efecto estacional.

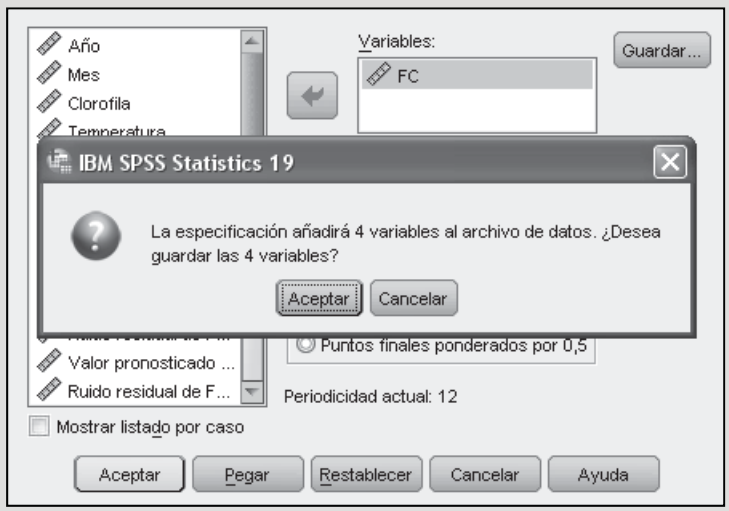
Puede ser de interés conocer si el estado reproductivo en verano responde únicamente a esa variación estacional esperada o si, más allá de la estacionalidad, el estado reproductivo está aumentando o disminuyendo en los diferentes años. Podemos comparar el valor de *FC* en julio de un año con los valores de otros años, y así comprobamos si el valor de *FC* está cambiando a corto plazo, es decir, no solo que sea alto en verano, sino que pueda estar cambiando de año en año.

Queremos aplicar algún método de análisis que nos permita comparar al mismo tiempo todos los valores de la serie. ¿Existe algún modo de eliminar el efecto estacional para poder ver y analizar los restantes efectos? El procedimiento adecuado es la desestacionalización, o corrección de la serie para eliminar la componente estacional. Para ello vamos a «Analizar» ⇨ «Predicciones» ⇨ «Descomposición estacional...».

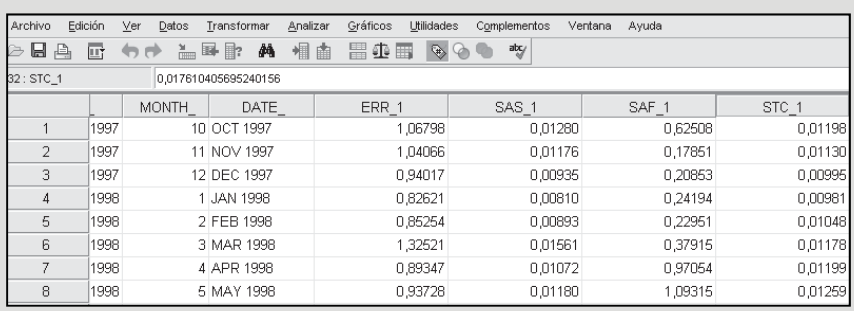


Cuadro XII.1. *(Continuación)* [SPSS]

Paso 27. En «Variables» se incorpora *FC* y al pulsar «Aceptar» aparece una ventana donde nos informa de que se van a crear varias variables y aceptamos de nuevo.



En el archivo de datos aparecen nuevas variables como la serie desestacionalizada «SAS_1», la tendencia/ciclo «STC_1», el efecto estacional «SAF_1» y el residuo «ERR_1».


Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Complementos Ventana Ayuda

32 : STC_1 0,017610405695240156

		MONTH_	DATE_	ERR_1	SAS_1	SAF_1	STC_1
1	1997	10	OCT 1997	1,06798	0,01280	0,62508	0,01198
2	1997	11	NOV 1997	1,04066	0,01176	0,17851	0,01130
3	1997	12	DEC 1997	0,94017	0,00935	0,20853	0,00995
4	1998	1	JAN 1998	0,82621	0,00810	0,24194	0,00981
5	1998	2	FEB 1998	0,85254	0,00893	0,22951	0,01048
6	1998	3	MAR 1998	1,32521	0,01561	0,37915	0,01178
7	1998	4	APR 1998	0,89347	0,01072	0,97054	0,01199
8	1998	5	MAY 1998	0,93728	0,01180	1,09315	0,01259

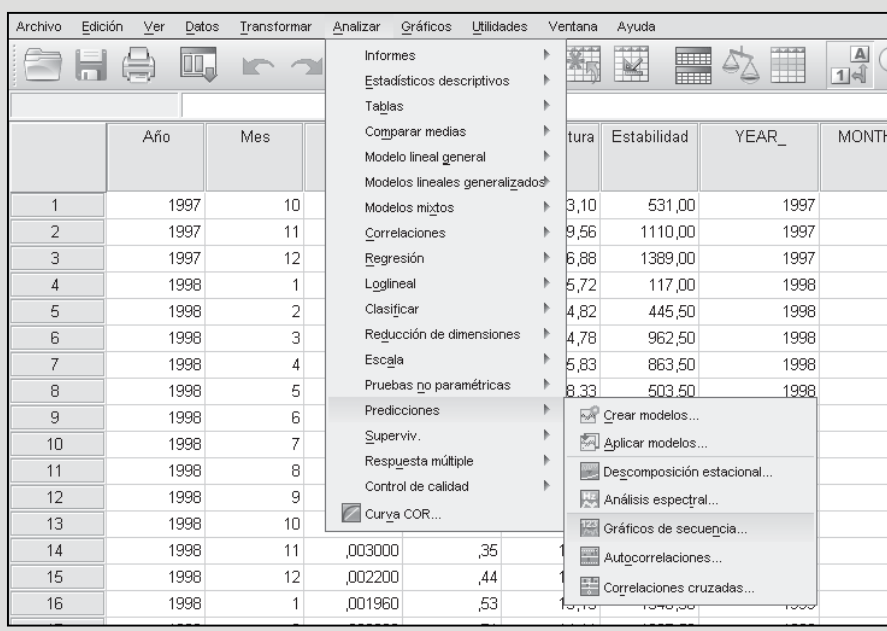
La nueva variable «SAF_1» contiene los índices de estacionalidad y podemos observar en el archivo de datos como se repiten los mismos doce valores todos los años. El valor más alto corresponde al mes de agosto, 2,48392, (la media anual se multiplica por ese valor en agosto, un incremento del 148,392%), y el más bajo a noviembre, 0,17851, una reducción del 82,149% sobre el valor medio anual en este mes sistemáticamente. Dividiendo cada valor de la serie original por su correspondiente índice estacional se obtiene la serie desestacionalizada «SAS_1», en la cual se ha eliminado por lo tanto el efecto de la estacionalidad. Esta nueva variable nos permite observar la evolución real de la serie –si aumenta o disminuye– más allá de los cambios puramente estacionales, que generalmente ocultan o enmascaran esta evolución.

Cuadro XII.1. *(Continuación)* SPSS

En la serie desestacionalizada permanecen todavía las componentes de tendencia, ciclo (si existe), y también las pequeñas oscilaciones residuales. Para suprimir estas últimas se realiza una operación de "suavizado" de la serie que consiste en sustituir cada valor –en la serie desestacionalizada– por la media de los doce valores que incluyen el actual, los seis anteriores, y los cinco siguientes (media móvil de orden 12). Así se obtiene la tendencia/ciclo «STC_1», que permite ver mejor la evolución de la variable al suprimir las pequeñas oscilaciones ocasionales.

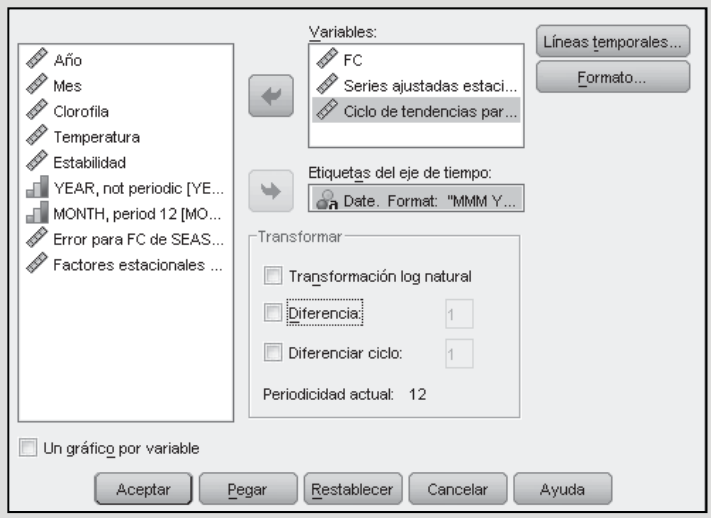
El cociente entre los valores de la serie desestacionalizada y los de tendencia/ciclo es el residuo «ERR_1», que en general tomará valores próximos a 1. Un valor de 1,06798, como el de octubre de 1998, indica que en ese mes el residuo es el 6,798% del valor de tendencia. Porcentajes pequeños indican que el modelo se ajusta razonablemente bien a los datos.

Paso 28. A continuación se hace un grafico de secuencia en «Analizar» ⇨ «Predicciones» ⇨ «Gráficos de secuencia...».

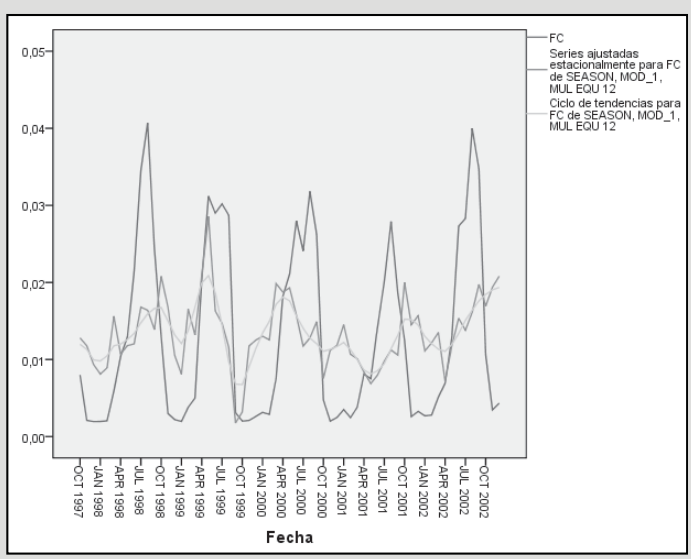


Paso 29. Representamos a continuación las nuevas variables. No debemos representar en el mismo gráfico las variables «Factores estacionales» «SAF_1» y el residuo o «Error para FC» «ERR_1», que están en una escala diferente a las restantes, Serie ajustada «SAS_1» y «Tendencia/Ciclo» «STC_1», estas dos en la misma escala de la variable original *FC*. Por ello, representamos «FC», «SAS_1» y «STC_1» mediante un gráfico de secuencia, con la variable de fecha para etiquetar el eje horizontal.

Cuadro XII.1. *(Continuación)* SPSS



Paso 30. Observamos, en color verde, la serie desestacionalizada, que podemos comparar con la original en azul. En color dorado la línea de tendencia/ciclo, que suaviza las pequeñas oscilaciones de apariencia aleatoria o residuales. En estas dos variables ha desaparecido el efecto estacional, por lo que podemos analizar la evolución. El valor más alto corresponde a mayo de 2000, y el más bajo a octubre de 2000, seis meses más tarde; quizás podamos encontrar las causas de estas variaciones estudiando la información complementaria, mediante otras variables relacionadas, sabiendo que estas causas son distintas de la estacionalidad. Si la variable tuviese una tendencia creciente o decreciente –no es así en nuestro ejemplo– se apreciaría claramente en la serie desestacionalizada o mejor aún en la serie de tendencia/ciclo.



XII.2. Ciclos

El ciclo es una componente habitual en las series temporales. Su forma típica es sinusoidal, con una fase de crecimiento hasta alcanzar el valor máximo, a la que sigue una fase de decrecimiento hasta alcanzar su valor mínimo, repitiéndose este esquema indefinidamente. A diferencia de las variaciones estacionales, no siempre es regular, y la duración de las distintas fases a menudo es cambiante, lo que dificulta su análisis. Está presente en prácticamente todas las series económicas, sin que se conozcan muy bien las razones de su aparición o las causas profundas de la existencia de ciclos.

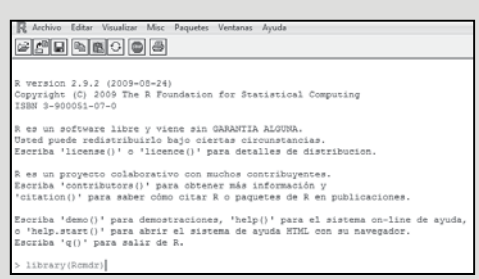
El análisis espectral es un procedimiento estadístico que puede ser utilizado para identificar comportamientos periódicos en una serie temporal, y para estudiar la existencia de ciclos o variaciones regulares cuya longitud (período) es generalmente superior a un año. El procedimiento descompone el conjunto de todas las observaciones de la serie en elementos (componentes) periódicos de diferentes frecuencias, y permite identificar las más relevantes. Es particularmente útil para detectar estructuras cíclicas complejas, es decir cuando existen varios ciclos superpuestos de diferente período, ya que cuando existe un único patrón cíclico puede ser suficiente una representación gráfica (gráfico secuencial) para identificarlo.

Cuadro XII.2. Análisis de ciclos

Aplicación con R.

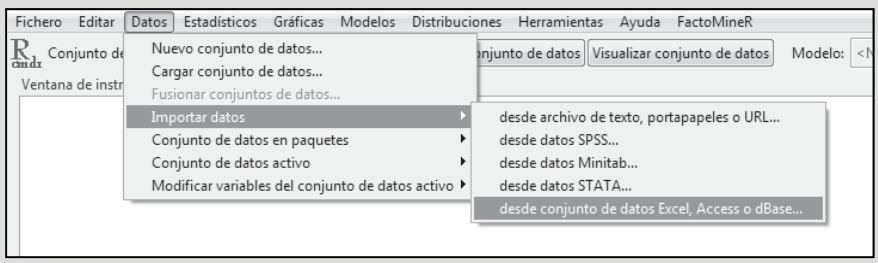
EJEMPLO. El ángulo que forman el sol y la luna es uno de los principales responsables de los cambios que se producen en la mareas. Cuando el sol y la luna están alineados, es decir, el ángulo es 0 (luna llena y luna nueva), las diferencias en altura entre pleamar y bajamar son máximas. Por el contrario, cuando el sol y la luna forman un ángulo de 90º (cuarto creciente y cuarto menguante) las diferencias en altura entre pleamar y bajamar son mínimas. En el archivo **Cuadro.XII.2.xls** están los ángulos entre el sol y la luna y la diferencia en metros entre pleamar y bajamar para 42º 8´Norte y 14º 13´Oeste del 1/1/2004 al 14/3/2004. El primer objetivo es determinar el ciclo de los cambios en el ángulo entre el sol y la luna. El segundo objetivo es demostrar la dependencia de las diferencias entre pleamar y bajamar de la posición entre el sol y la luna.

Paso 1. Iniciamos R y elegimos el directorio de trabajo. Son necesarios los paquetes "Rcmdr" y "seewave". Se introduce la instrucción «library(Rcmdr)».

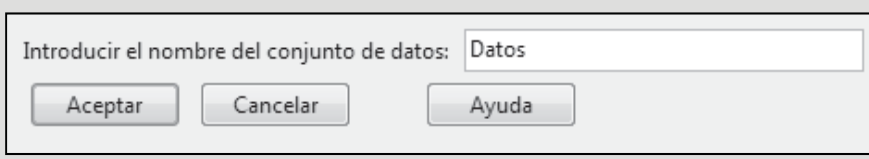

Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> library(Rcmdr)

Cuadro XII.2. *(Continuación)*[R]

Paso 2. Se selecciona «Datos» ⇨ «Importar datos» ⇨ «desde conjunto de datos Excel, Access o dBase».



Paso 3. Aparece un pequeño cuadro de diálogo para introducir el nombre del conjunto de datos, en el que dejamos el nombre por defecto «Datos» y pulsamos «Aceptar». Posteriormente aparecen los archivos del directorio de trabajo, en el que elegimos **Cuadro XII.2.xls** y pulsamos «Abrir».

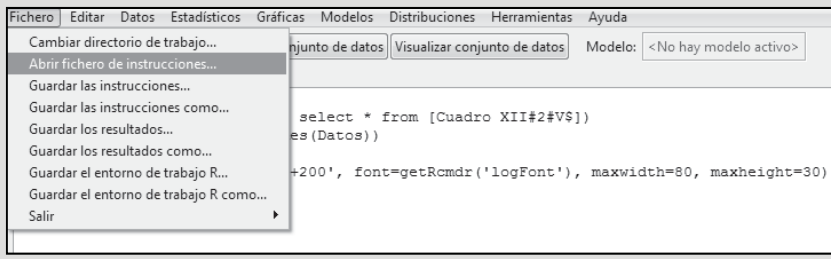


Paso 4. En R Commander pulsamos en «Visualizar conjunto de datos». Es importante inspeccionar los datos y comprobar que no existan líneas, lo cual a veces ocurre al final del archivo, que tengan celdas con el código <NA>. Si existen estas líneas, hay que eliminarlas.

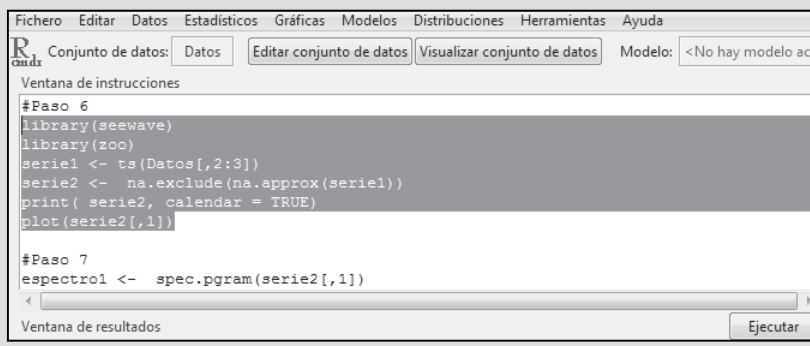
	Fecha	Angulo.Sol.Luna	Diferencia.Pleamar.Bajamar
1	2004-01-01	53.33	1.47
2	2004-01-02	38.40	1.47
3	2004-01-03	26.20	1.54
4	2004-01-04	16.65	1.70
5	2004-01-05	8.78	1.92
6	2004-01-06	3.00	2.20
7	2004-01-07	0.20	2.37
8	2004-01-08	1.20	2.55
9	2004-01-09	5.63	2.63
10	2004-01-10	13.37	2.62
11	2004-01-11	23.18	2.61
12	2004-01-12	36.00	2.44
13	2004-01-13	52.20	2.31
14	2004-01-14	70.20	2.09
15	2004-01-15	88.20	1.93
16	2004-01-16	68.40	1.80
17	2004-01-17	48.00	1.88
18	2004-01-18	29.40	1.95
19	2004-01-19	15.30	2.27
20	2004-01-20	5.10	2.64
21	2004-01-21	0.60	3.00
22	2004-01-22	1.80	3.36
23	2004-01-23	6.90	3.44
24	2004-01-24	17.49	3.35
25	2004-01-25	31.20	2.98
26	2004-01-26	46.50	2.75

Cuadro XII.2. *(Continuación)*[R]

Paso 5. A continuación se carga el archivo de instrucciones **Cuadro XII.2.R** seleccionando «Fichero» ⇨ «Abrir fichero de instrucciones…» y no se guarda la salida.



Paso 6. Se seleccionan las primeras 6 líneas y se pulsa «Ejecutar». Los resultados y cada una de las instrucciones se explican a continuación.



Las primeras instrucciones cargan los paquetes necesarios.

```
library(seewave)
library(zoo)
```

Se construye la serie temporal. Esta operación asigna a los datos el formato adecuado en R para su tratamiento estadístico posterior. Se define la serie, que llamaremos «serie1», la cual incluye las columnas 2 a 3 de los datos. No indicamos periodicidad, ya que la serie de tiempo es simplemente una secuencia de días consecutivos. La primera variable es el ángulo entre el sol y la luna, y la segunda la diferencia entre pleamar y bajamar.

```
serie1 <- ts(Datos[,2:3]) # convierte los datos en serie temporal
```

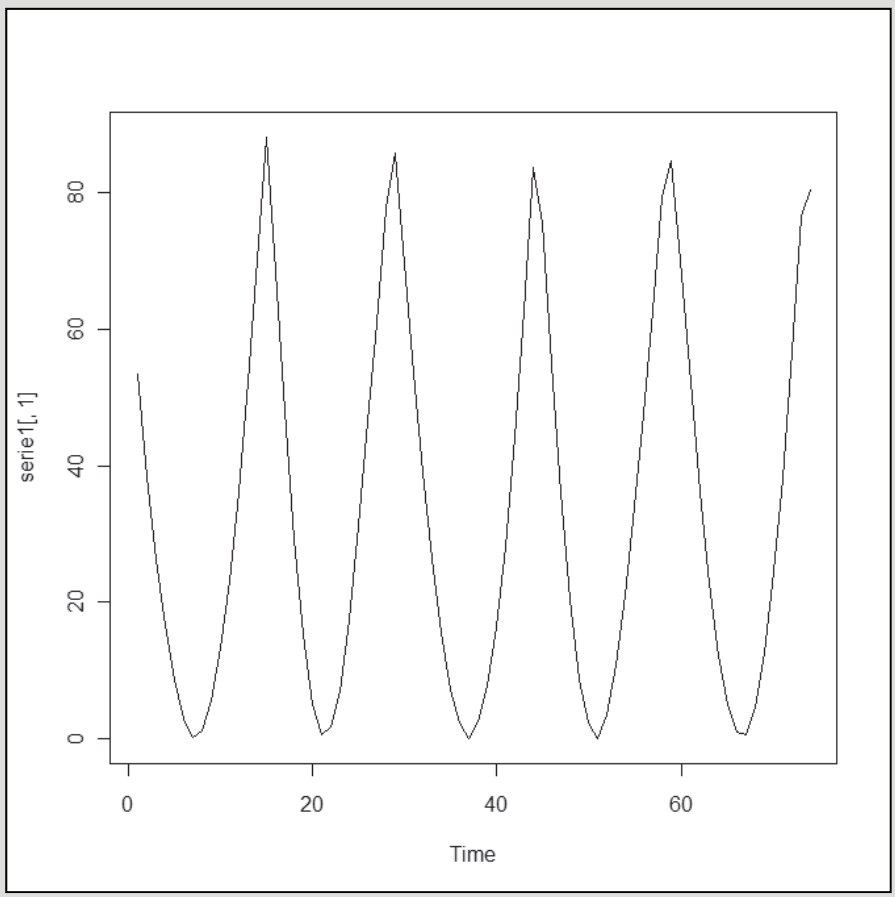
En el supuesto de que falten valores en la serie (que no es el caso de este ejemplo) se crea una nueva serie, en la cual la función «na.approx» sustituye los valores perdidos por valores interpolados.

```
serie2 <-  na.exclude(na.approx(serie1)) # sustituye los valores que faltan por valores interpolados
```

Cuadro XII.2. *(Continuación)*[R]

Imprimimos y representamos gráficamente la variable de interés, "Angulo.Sol.Luna", la primera de la serie múltiple creada.

print(serie2, calendar = TRUE) *# imprime la serie completa*
plot(serie2[,1]) *# gráfico de secuencia de la primera variable (Angulo.Sol.Luna)*

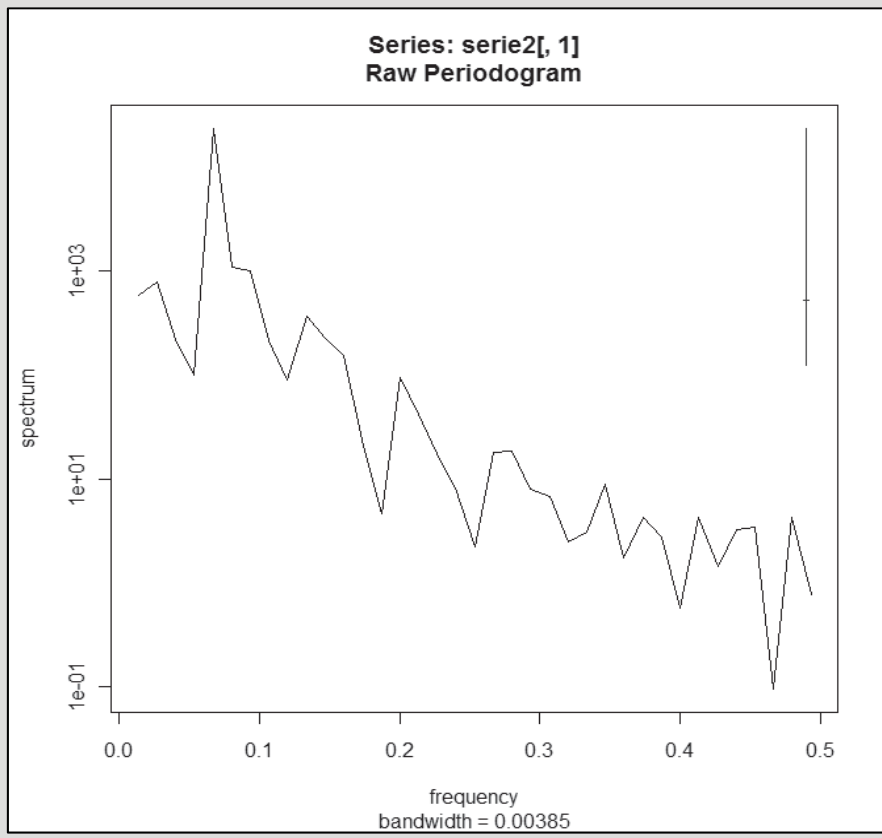


El gráfico muestra una clara variación cíclica, con 5 oscilaciones en el conjunto de los 74 días, lo que corresponde a un período de aproximadamente 15 días. Podemos profundizar un poco más mediante un análisis espectral, que estudia el comportamiento o la respuesta de la serie a cada frecuencia diferente.

Paso 7. A continuación construimos el espectrograma. Para ello podemos usar el periodograma mediante la función de R «spec.pgram» aplicada a la primera de nuestras variables, «serie2[,1]». Además del gráfico, los resultados quedan almacenados en una variable que llamaremos «espectro1».

Cuadro XII.2. *(Continuación)*[R]

espectro1 <- spec.pgram(serie2[,1]) *# periodograma*



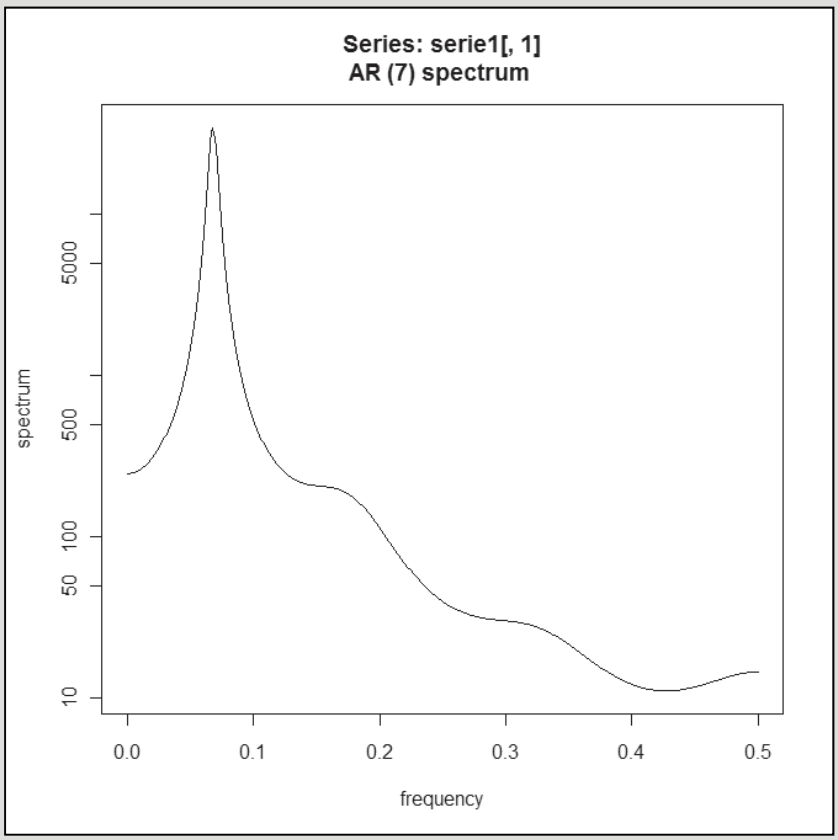
En el gráfico se puede observar claramente un pico que destaca sobre el resto de valores. Ese máximo indica la frecuencia del ciclo que hemos observado anteriormente en el gráfico de secuencia. Si hubiese más ciclos superpuestos, encontraríamos otros picos en el espectrograma mostrando las frecuencias asociadas a ellos.

Paso 8. El espectograma también se puede representar mediante el gráfico de densidad espectral, el cual es simplemente una versión suavizada del periodograma, eliminando mediante filtrado el ruido formado por las variaciones aleatorias, o en general cualquier variación de alta frecuencia. Además del gráfico, en este caso los resultados quedan almacenados en una variable que llamaremos «espectro2».

espectro2 <- spec.ar(serie2[,1]) *# diagrama espectral (transformada de Fourier)*

El efecto de suavizar los datos nos permite ver de forma más clara el pico que se observaba en el periodograma.

Cuadro XII.2. *(Continuación)*[R]



Entre ambos gráficos, se puede representar y utilizar el que presente los resultados más claros. Es posible que cuando falten valores en la serie, sea más conveniente utilizar el primer gráfico, es decir, el periodograma. Sin embargo, como en este caso los resultados son muy parecidos utilizaremos el segundo gráfico, el gráfico de densidad espectral.

Paso 9. Para determinar con mayor precisión la duración del ciclo examinamos los resultados numéricos, almacenados en la variable «espectro2»: «espectro2$spec», que muestra los valores de densidad espectral representados en el espectrograma, y «espectro2$freq», que indica la frecuencia asociada a cada uno de ellos (en R el símbolo $ se emplea para identificar los distintos elementos o componentes de una variable estructurada).

```
Espectro2$spec # imprime la densidad espectral
1/espectro2$freq # imprime el inverso de la frecuencia, es decir el período
```

En la lista de valores buscamos el máximo (34784,98728), que corresponde al numero 69.

Cuadro XII.2. *(Continuación)*[R]

```
[63,]  9656.30846
[64,] 12412.86863
[65,] 16199.20838
[66,] 21227.55585
[67,] 27230.19583
[68,] 32680.34907
[69,] 34784.98728
[70,] 32078.25422
[71,] 26404.95388
[72,] 20480.08834
[73,] 15620.18375
[74,] 11988.16320
[75,]  9347.94910
```

El período, o inverso de la frecuencia, viene expresado en fracción de día (que es la unidad de tiempo de nuestros datos)

```
90.727273  83.166667  76.769231  71.285714  66.533333
34.413793  33.266667  32.193548  31.187500  30.242424
21.234043  20.791667  20.367347  19.960000  19.568627
15.353846  15.121212  14.895522  14.676471  14.463768
12.024096  11.880952  11.741176  11.604651  11.471264
 9.881188   9.784314   9.689320   9.596154   9.504762
```

El valor número 69 es 14,676 días. Este es el período medio estimado del ciclo observado.

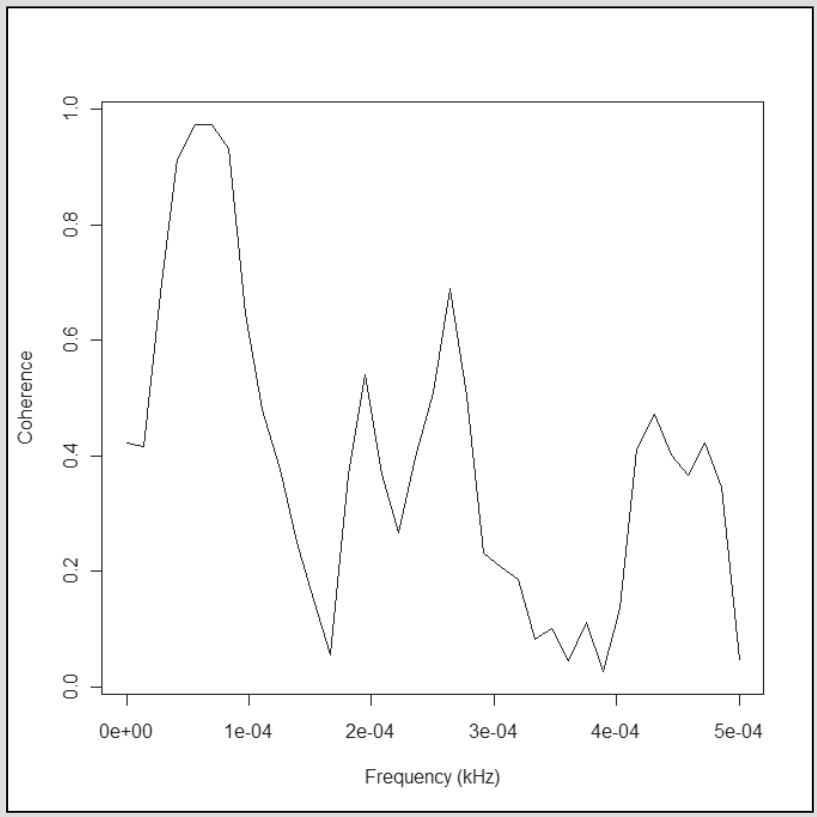
Paso 10. El segundo objetivo es averiguar si la diferencia entre pleamar y bajamar está relacionada con el ángulo entre el sol y la luna. Para ello debemos construir el espectro cruzado de ambas variables.

```
coherencia <- coh(serie2[,1],serie2[,2]) # construye el diagrama de coherencia del espectro cruzado
coherencia # escribe los valores de frecuencia y coherencia representados en el gráfico
```

El espectro cruzado se interpreta habitualmente a partir de algunas variables transformadas. Una de las más importantes es la coherencia cuadrada, representada en el diagrama siguiente, que toma valores entre 0 y 1. Un valor próximo a 0 indica que ambas variables no están relacionadas, y próximo a 1 indica una relación prácticamente lineal entre ambas. La escala horizontal muestra la frecuencia en kHz, ya que es la unidad más útil para el estudio de señales en el campo de la electrónica en el que esta técnica es usualmente aplicada. Deben multiplicarse por mil los valores para interpretarlos directamente como fracción de la unidad de tiempo empleada.

Cuadro XII.2. *(Continuación)*[R]

Como se puede observar en el gráfico existe un valor próximo a 1 que destaca claramente, mostrando para la frecuencia correspondiente una acusada relación cíclica entre ambas variables.



El valor máximo es 0,9739, muy próximo a 1, y corresponde a la frecuencia 0,069444 (recordemos que está dividido por 1000). El inverso de la frecuencia, el período 1/0,06944 = 14,4 días es la duración del ciclo compartido entre ambas variables.

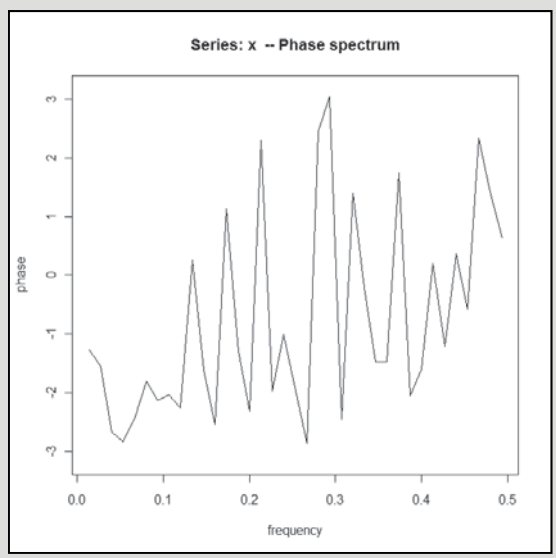
```
> coherencia
                    X
[1,] 0.000000e+00 0.42302897
[2,] 1.388889e-05 0.41656461
[3,] 2.777778e-05 0.68090727
[4,] 4.166667e-05 0.91260767
[5,] 5.555556e-05 0.97232042
[6,] 6.944444e-05 0.97390973
[7,] 8.333333e-05 0.93169827
[8,] 9.722222e-05 0.64498026
[9,] 1.111111e-04 0.47993779
```

Cuadro XII.2. *(Continuación)*[R]

Paso 11. Aún existiendo un ciclo coincidente en duración para las dos variables, es posible que exista un desfase o retardo entre ambos. Para averiguarlo construimos el diagrama de fase:

```
espectro <-  spectrum(serie2) # calcula el espectro cruzado
plot(espectro, plot.type = "phase") # grafico de ángulo de desfase para cada período
espectro$phase # imprime el ángulo de desfase para cada período
```

En el diagrama se observan oscilaciones irregulares –de apariencia aleatoria– en torno al cero, con amplitud entre -3 y +3 días, para la mayoría de las frecuencias, pero para la frecuencia coincidente que hemos detectado, por debajo de la frecuencia 0,1, existen varios valores que parecen estabilizarse alrededor de -2.



La siguiente tabla muestra los valores de desfase representados en el gráfico. Hay que fijarse en el sexto valor (véase en los datos de coherencia que el máximo era 0,9739 para el sexto valor), el cual es -1,8. Al ser negativo el desfase significa que el efecto del ángulo entre el sol y la luna se produce 1,8 días después en las mareas.

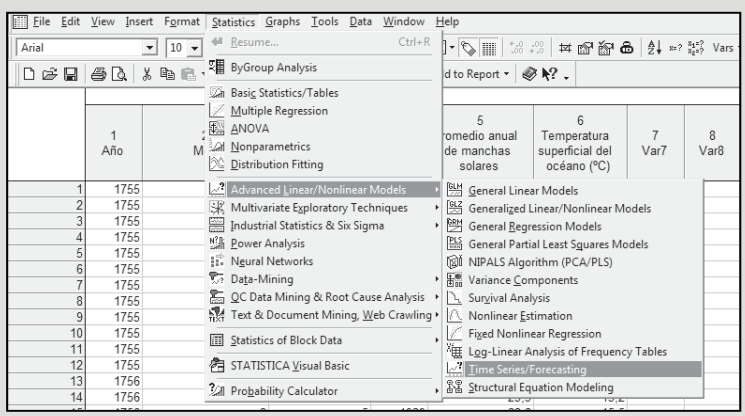
```
> espectro$phase
             [,1]
[1,] -1.2704888
[2,] -1.5522780
[3,] -2.6771422
[4,] -2.8300287
[5,] -2.4095503
[6,] -1.8032338
[7,] -2.1332456
```

Cuadro XII.2. *(Continuación)*[STATISTICA]

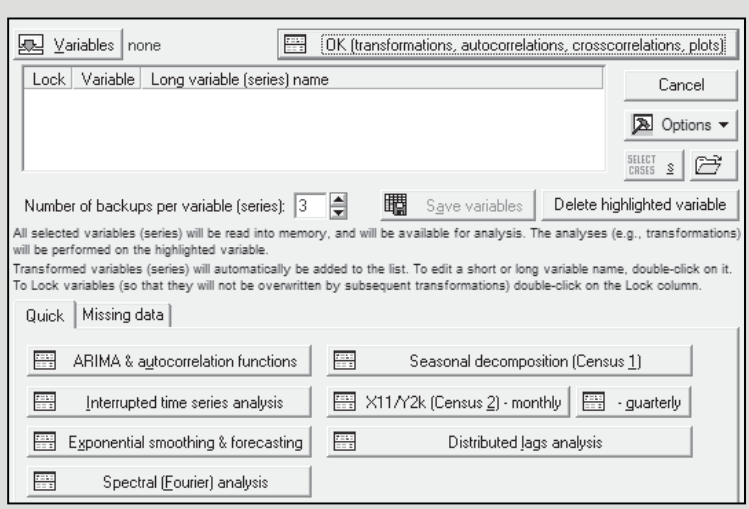
Aplicación con STATISTICA.

EJEMPLO. Los datos están en el archivo **Cuadro XII.2.sta**, que son datos del número de manchas solares mensuales desde el año 1755 hasta el 2007, la temperatura superficial del océano a 42ºN 10ºO desde el año 1925 al 1996 y el promedio anual de manchas solares para ese mismo periodo. El primer objetivo es determinar si existe un ciclo en la serie temporal de manchas solares. Por otro lado, la importancia del seguimiento de las manchas solares radica en que se ha observado que pueden tener una importancia grande sobre el clima en la tierra, ya que parece ser que la temperatura de la tierra aumenta a medida que el número de manchas solares es mayor. Por tanto, el segundo objetivo es determinar si existe relación entre las manchas solares y la temperatura en el océano.

Paso 1. En primer lugar se accede a «Statistics» ⇨ «Advanced/Nonlinear Models» ⇨ «Time Series/Forecasting».

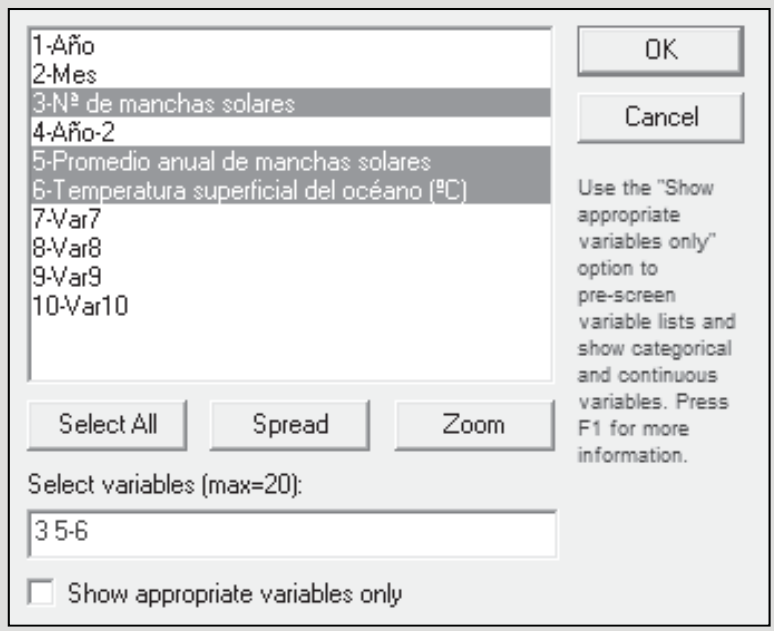


Paso 2. En la siguiente ventana se pulsa en «Variables».

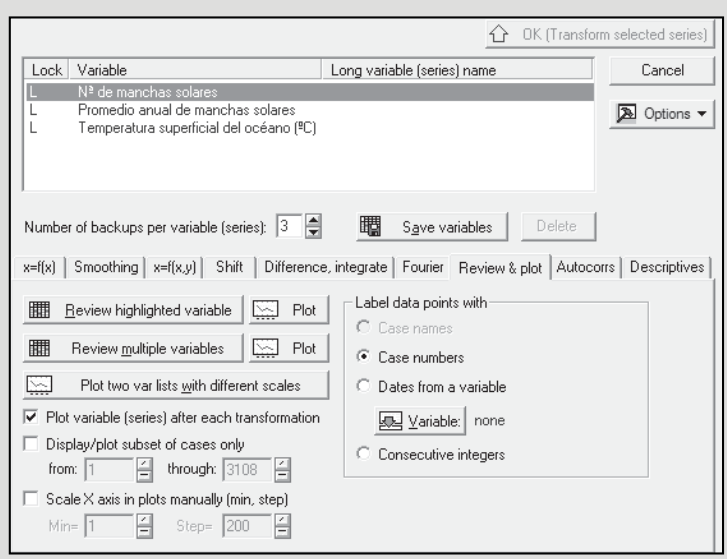

Variables none
OK (transformations, autocorrelations, crosscorrelations, plots)
Lock Variable Long variable (series) name
Cancel
Options
Number of backups per variable (series): 3
Save variables
Delete highlighted variable
All selected variables (series) will be read into memory, and will be available for analysis. The analyses (e.g., transformations) will be performed on the highlighted variable.
Transformed variables (series) will automatically be added to the list. To edit a short or long variable name, double-click on it.
To Lock variables (so that they will not be overwritten by subsequent transformations) double-click on the Lock column.
Quick Missing data
ARIMA & autocorrelation functions
Seasonal decomposition (Census 1)
Interrupted time series analysis
X11/Y2k (Census 2) - monthly
- quarterly
Exponential smoothing & forecasting
Distributed lags analysis
Spectral (Fourier) analysis

Cuadro XII.2. *(Continuación)*[STATISTICA]

Paso 3. En la pantalla que aparece se seleccionan las variables «Nº de manchas solares», «Promedio anual de manchas solares» y «Temperatura superficial del océano (ºC)».

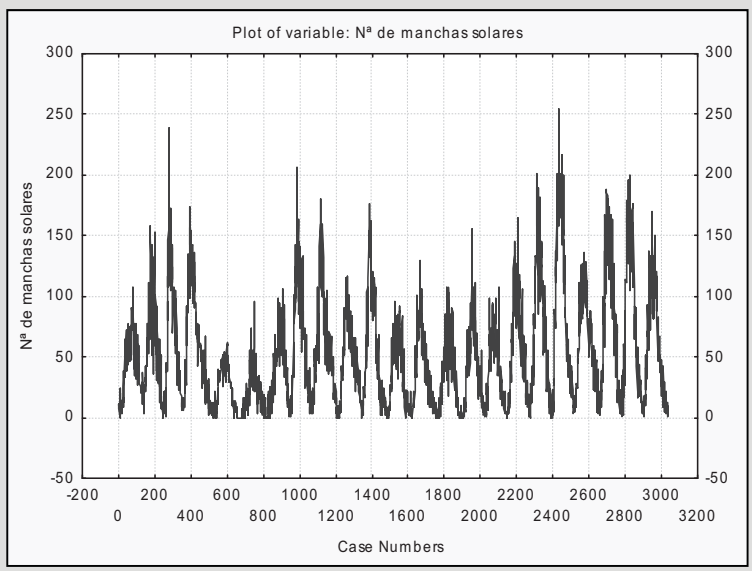


Paso 4. Para obtener una representación de los datos, en la ventana del paso 2 se pulsa en «OK (transformations, autocorrelations, crosscorrelations, plots)» y en la pestaña «Review & plot», de la ventana que aparece, pulsamos en «Plot» de «Review highlighted variable». Hay que asegurarse de que en la ventana superior esté seleccionado «Nº de manchas solares».

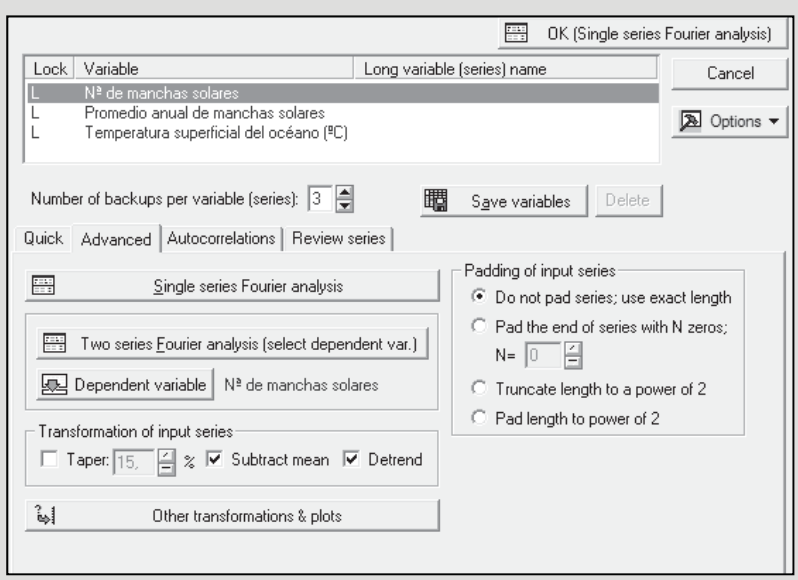

OK (Transform selected series)
Lock
Variable
Long variable (series) name
L Nª de manchas solares
L Promedio anual de manchas solares
L Temperatura superficial del océano (ºC)
Cancel
Options
Number of backups per variable (series): 3
Save variables
Delete
x=f(x) | Smoothing | x=f(x,y) | Shift | Difference, integrate | Fourier | Review & plot | Autocorrs | Descriptives
Review highlighted variable
Plot
Review multiple variables
Plot
Plot two var lists with different scales
Plot variable (series) after each transformation
Display/plot subset of cases only
from: 1 through: 3108
Scale X axis in plots manually (min, step)
Min= 1 Step= 200
Label data points with
Case names
Case numbers
Dates from a variable
Variable: none
Consecutive integers

Cuadro XII.2. *(Continuación)*^{STATISTICA}

Paso 5. La gráfica que se obtiene muestra una clara variación cíclica, con 24 oscilaciones en el conjunto de 259 años, lo que corresponde a un período de aproximadamente 11 años. Para profundizar un poco más realizaremos un análisis espectral.

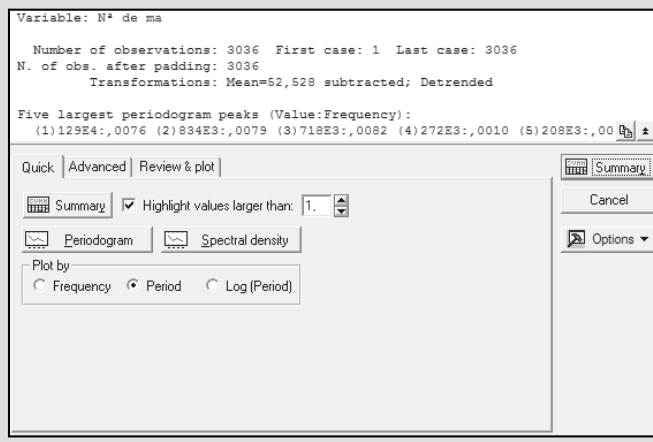


Paso 6. Se pulsa «Cancel» en la ventana del paso 4 y aparece la ventana del paso 2 donde se selecciona «Spectral (Fourier) analysis». Se obtiene la siguiente ventana donde, asegurándonos de que esté seleccionado «Nº de manchas solares», se pulsa en «Single Series Fourier analysis».

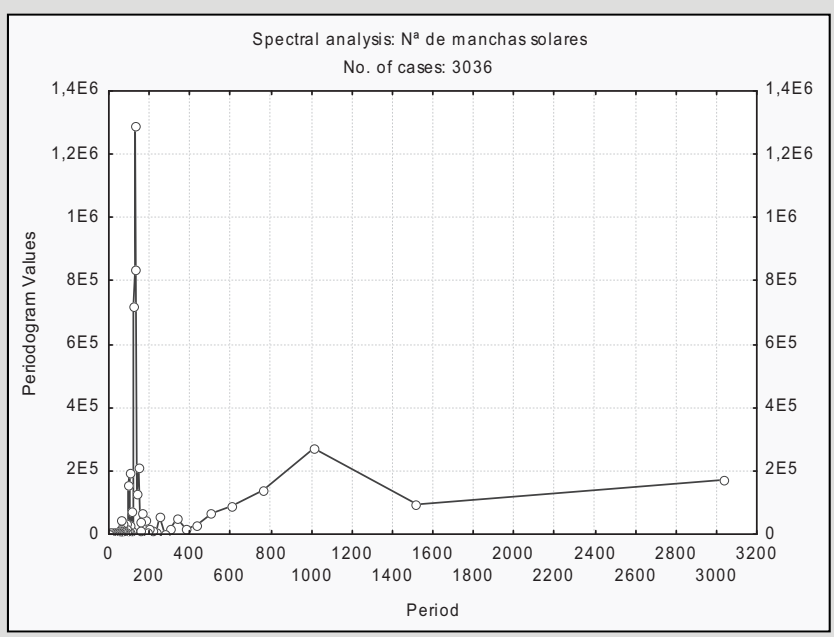


Cuadro XII.2. *(Continuación)*[STATISTICA]

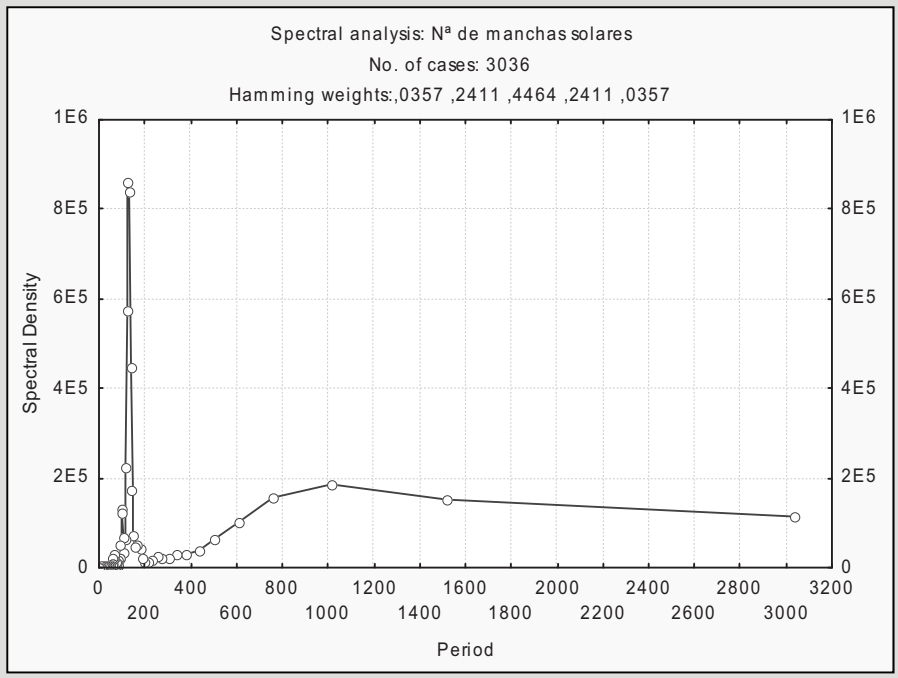
Paso 7. En la pestaña «Quick» en «Plot by» se selecciona «Period». Es posible elegir entre dos tipos de gráficos «Periodogram» y «Spectral density». El gráfico de densidad espectral es simplemente una versión suavizada del periodograma, eliminando mediante filtrado el ruido formado por las variaciones aleatorias, o en general cualquier variación de alta frecuencia. Se puede representar y utilizar el que presente los resultados más claros.



Paso 8. Ambos gráficos muestran un pico muy claro y otro también importante.



Cuadro XII.2. *(Continuación)* *STATISTICA*



Paso 9. Para conocer el periodo de ambos ciclos en la ventana del paso 7 pulsamos en «Summary». En el periodograma, el primer pico se produce a 132 meses, y en el gráfico de densidad a 126,5 meses. El promedio de ambos es 129,25 meses. Este patrón cíclico se conoce como ciclo de las manchas solares y es el que se apreciaba claramente en la gráfica del paso 5, que ahora sabemos que tiene un periodo medio de 10,77 años (129,25 meses). Es importante resaltar el carácter de valor promedio de este dato, ya que los ciclos pueden variar en el tiempo entre aproximadamente unos 8 a 15 años.

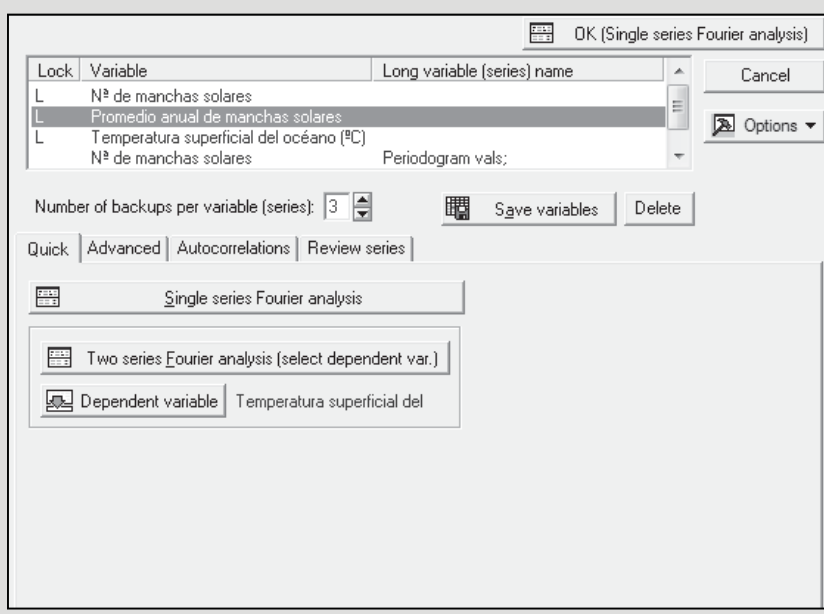
	Spectral analysis: Nª de manchas solares (Cuadro XII.2.sta) No. of cases: 3036					
	Frequency	Period	Cosine Coeffs	Sine Coeffs	Periodogram	Density
20	0,006588	151,800	2,4162	-1,1292	10798	71902,2
21	0,006917	144,571	1,6019	-11,5947	207972	173969,0
22	0,007246	138,000	1,8859	8,9977	128294	448895,5
23	0,007576	132,000	-27,7416	-9,0068	1291394	841530,4
24	0,007905	126,500	-10,0394	21,1788	833887	861603,4
25	0,008235	121,440	-0,9529	21,7235	717736	572951,0
26	0,008564	116,769	1,1256	-2,4973	11391	225825,3
27	0,008893	112,444	-4,8118	-5,0534	73912	63111,6

Cuadro XII.2. *(Continuación)*[STATISTICA]

El segundo pico, tanto en el periodograma cómo en el gráfico de densidad espectral, es a 1012 meses, es decir, que tiene un periodo promedio de 84,3 años y se conoce como ciclo de Gleissberg. Este ciclo se piensa que oscila entre los 80 y 100 años aproximadamente. Sin embargo, como ya se mostró anteriormente en las gráficas del paso 5, no es un ciclo tan claro como el de las manchas solares.

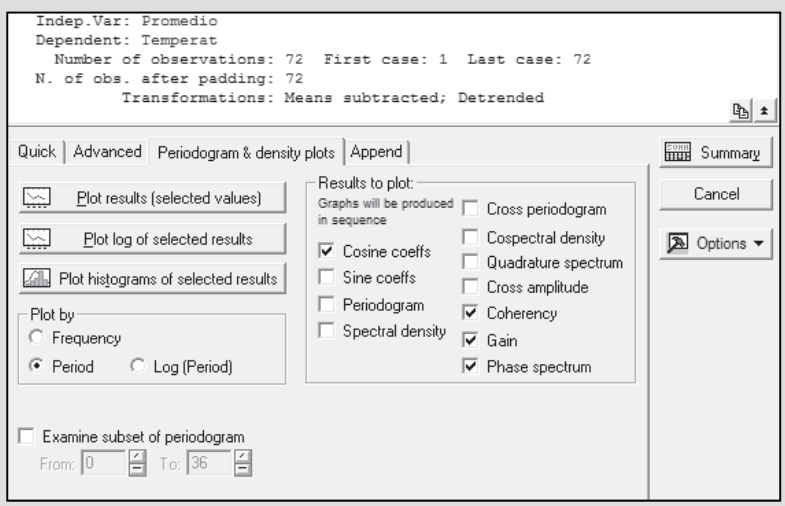
	Spectral analysis: Nª de manchas solares (Cuadro XII.2.sta) No. of cases: 3036					
	Frequency	Period	Cosine Coeffs	Sine Coeffs	Periodogram	Density
1	0,000329	3036,000	10,5102	-1,4421	170843	114929,7
2	0,000659	1518,000	1,0525	-7,8301	94750	154056,3
3	0,000988	1012,000	-2,1027	13,2218	272082	187142,3
4	0,001318	759,000	-7,5516	5,9025	139452	154668,8
5	0,001647	607,200	-5,6184	-5,0903	87251	99483,7
6	0,001976	506,000	-2,4823	-6,1799	67327	63295,3
7	0,002306	433,714	-3,7453	1,9491	27061	37849,3
8	0,002635	379,500	-2,8306	2,2165	19621	29762,0

Paso 10. Para abordar el segundo objetivo del trabajo, es decir, determinar si el número de manchas solares afecta a la temperatura del océano, se pulsa «Cancel» en la ventana del paso 7 y aparece la siguiente ventana, donde pulsamos en «Dependent variable» y se selecciona «Temperatura superficial del océano (ºC)». Nos aseguramos de que esté seleccionado «Promedio anual de manchas solares» en la ventana superior y pulsamos en «Two series Fourier (select dependent var.)».

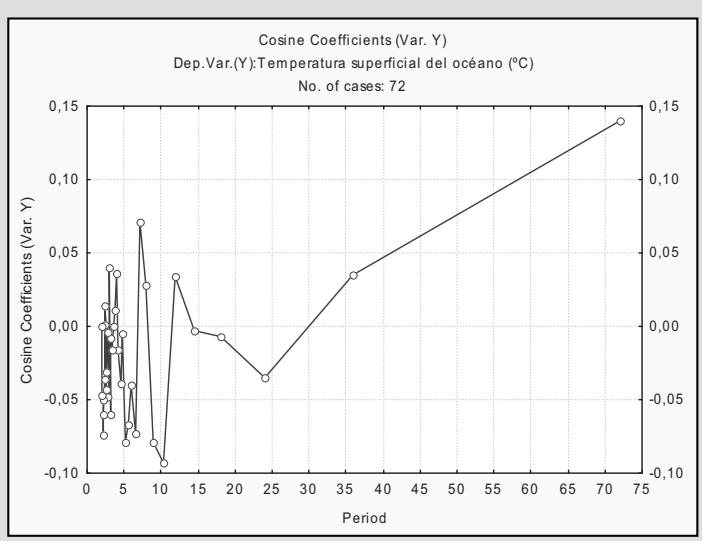


Cuadro XII.2. *(Continuación)*[STATISTICA]

Paso 11. El espectro cruzado descompone la serie temporal conjunta (de dos variables) en una suma de componentes de distintas frecuencias, y es la transformación de Fourier de la función de covarianza cruzada, que mide la asociación o relación entre ambas variables. Tiene forma de número complejo: su parte real se llama coespectro (Cross-p. real) y su parte imaginaria espectro cuadrado (Cross-p. imagin.), y es muy difícil de interpretar. Por ello se utilizan otras funciones derivadas de esta transformación, entre las cuales se usan más frecuentemente los coeficientes del coseno, Coherencia, Ganancia y el desplazamiento de fase. En la ventana que aparece, en la pestaña «Periodogram & density plots», se selecciona «Cosine coeffs», «Coherency», «Gain» y «Phase spectrum» y, por último, en «Plot by» se marca en «Period».



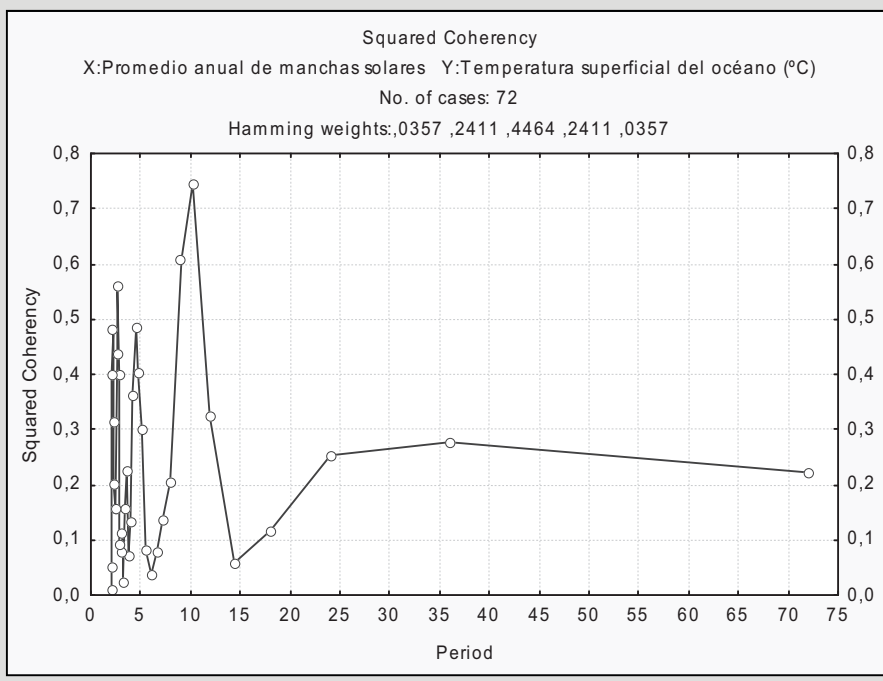
Paso 12. Al pulsar en «Plot results (selected values)» aparecen cada uno los gráficos de los parámetros seleccionados.



Cuadro XII.2. *(Continuación)*[STATISTICA]

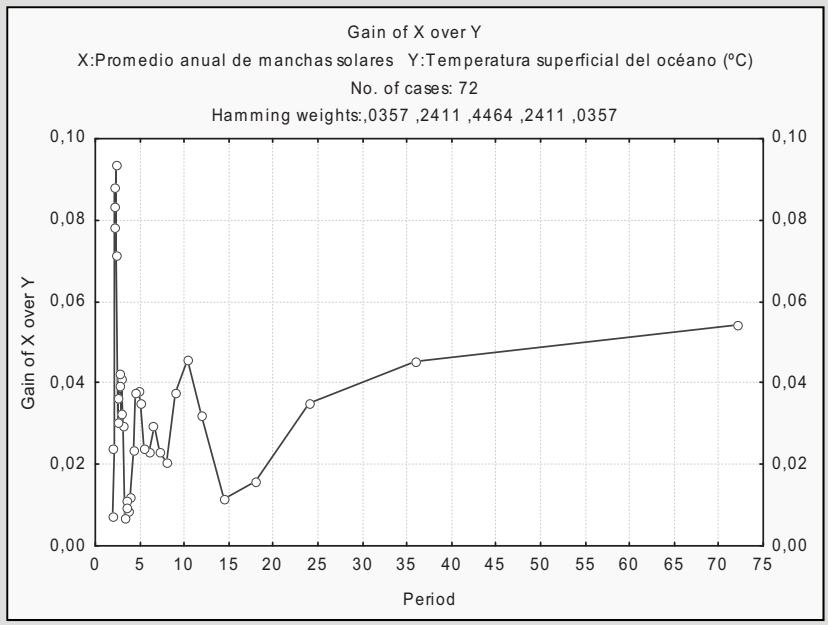
En la grafica anterior se muestra el coeficiente del coseno de la variable *Y*. Los coeficientes del coseno se pueden interpretar como coeficientes de regresión; indican en qué grado las respectivas funciones coseno están correlacionadas entre sí para cada frecuencia. Estos coeficientes se pueden calcular para las variables *X* e *Y*. En nuestro caso nos interesa la variable *Y*, es decir, «Temperatura superficial del océano (ºC)». En la gráfica se observa un valor negativo pequeño alrededor de los 10 años.

El coeficiente de Coherencia «Squared Coherency», que se muestra en la siguiente gráfica, mide el cuadrado de la correlación lineal entre dos variables para cada frecuencia. Un valor próximo a 1 indica que las dos variables están muy relacionadas para esa frecuencia (próximo a 0 indica lo contrario). De nuevo observamos una correlación entre el número de manchas solares y la temperatura del océano alrededor de los 10 años.

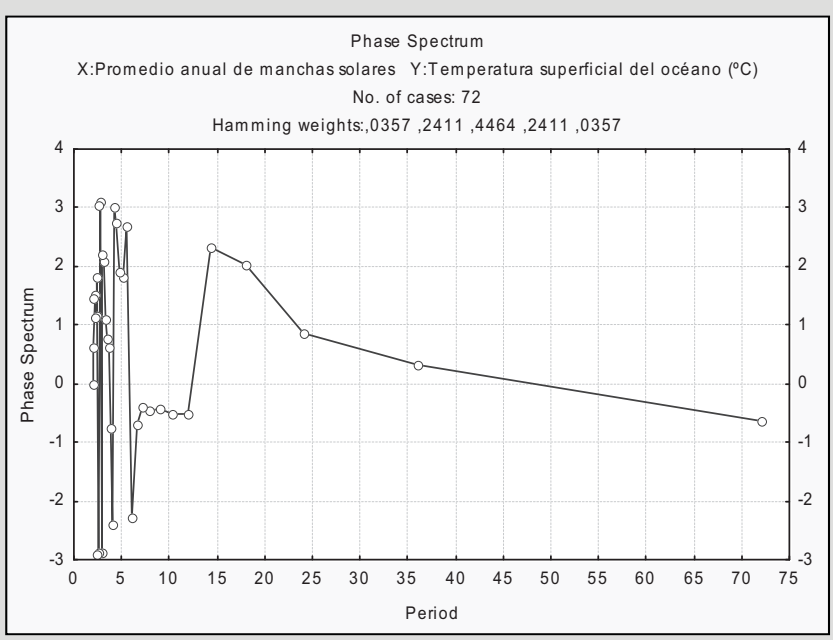


En la siguiente gráfica se muestra la Ganancia, que es el número por el que se multiplica la amplitud de una variable para reproducir la componente de amplitud de la otra variable para esa frecuencia (es como la pendiente en una regresión lineal); cuanto más grande, más relación hay. Se puede calcular la ganancia de *X* sobre *Y* o de *Y* sobre *X*. Lógicamente interesa el valor de *X* sobre *Y*, que es el que se muestra en la gráfica. De nuevo observamos valores altos alrededor de los 10 años.

Cuadro XII.2. *(Continuación)* *STATISTICA*


Gain of X over Y
X:Promedio anual de manchas solares Y:Temperatura superficial del océano (ºC)
No. of cases: 72
Hamming weights:,0357 ,2411 ,4464 ,2411 ,0357
Gain of X over Y
Period
0,00
0,02
0,04
0,06
0,08
0,10
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75

El Desplazamiento de fase «Phase spectrum» mide la diferencia (en este caso en años) o desfase entre ambas series para cada frecuencia. Se observa que alrededor de los 10 años, la diferencia es de aproximadamente medio año.



Cuadro XII.2. *(Continuación)*[STATISTICA]

Paso 13. Para poder ver los valores exactos en la ventana del paso 11 se pulsa en «Summary». Los resultados se muestran en la siguiente tabla.

	Bivariate Spectral Analysis; No. cases: 72 (Cuadro XII.2.sta) Indep.(X): Promedio anual de manchas solares Dep.(Y): Temperatura superficial del océano (ºC)					
	Frequncy	Period	Y Cosine	Squared Coherncy	Gain val X over Y	Phase Spectrum
0	0,000000		0,000000	0,063432	0,037359	0,00000
1	0,013889	72,00000	0,139680	0,221612	0,054273	-0,63268
2	0,027778	36,00000	0,034966	0,277751	0,045388	0,31278
3	0,041667	24,00000	-0,034635	0,254725	0,035143	0,85805
4	0,055556	18,00000	-0,006857	0,115750	0,015777	2,03675
5	0,069444	14,40000	-0,002587	0,056931	0,011362	2,32993
6	0,083333	12,00000	0,034534	0,326496	0,032043	-0,51405
7	0,097222	10,28571	-0,092555	0,744802	0,045737	-0,50994
8	0,111111	9,00000	-0,078497	0,607845	0,037829	-0,41872
9	0,125000	8,00000	0,028538	0,203896	0,020348	-0,46301
10	0,138889	7,20000	0,071176	0,135622	0,023216	-0,40125
11	0,152778	6,54545	-0,072656	0,080435	0,029375	-0,68316
12	0,166667	6,00000	-0,040092	0,036803	0,023080	-2,27498
13	0,180556	5,53846	-0,066695	0,082954	0,024129	2,67614
14	0,194444	5,14286	-0,079077	0,301508	0,035162	1,81066
15	0,208333	4,80000	-0,004862	0,403589	0,038041	1,92046
16	0,222222	4,50000	-0,038485	0,485852	0,037699	2,75276
17	0,236111	4,23529	-0,016129	0,361836	0,023499	3,00059
18	0,250000	4,00000	0,035695	0,133195	0,011967	-2,39624
19	0,263889	3,78947	0,010657	0,071600	0,008361	-0,75987
20	0,277778	3,60000	0,000393	0,226220	0,010924	0,62765
21	0,291667	3,42857	-0,015889	0,158042	0,009347	0,76474
22	0,305556	3,27273	-0,007510	0,022798	0,006851	1,10830
23	0,319444	3,13043	-0,060290	0,112607	0,029495	2,09142
24	0,333333	3,00000	0,040162	0,079645	0,040991	2,19798

Se observa que los valores absolutos más altos del coeficiente del coseno para el eje *Y* (-0,092), el coeficiente de Coherencia (0,74) y la Ganancia de *X* sobre *Y* (0,045) se dan a 10,28 años. En el análisis espectral de los ciclos de las manchas solares se observó que había un ciclo de manchas solares alrededor de los 10,77 años. Sin embargo, los valores no son significativos, ya que no están marcados en rojo. Por lo tanto, se puede concluir que no existe un efecto claro del número de manchas solares sobre la temperatura del océano, tan solo indicios de un efecto débil.

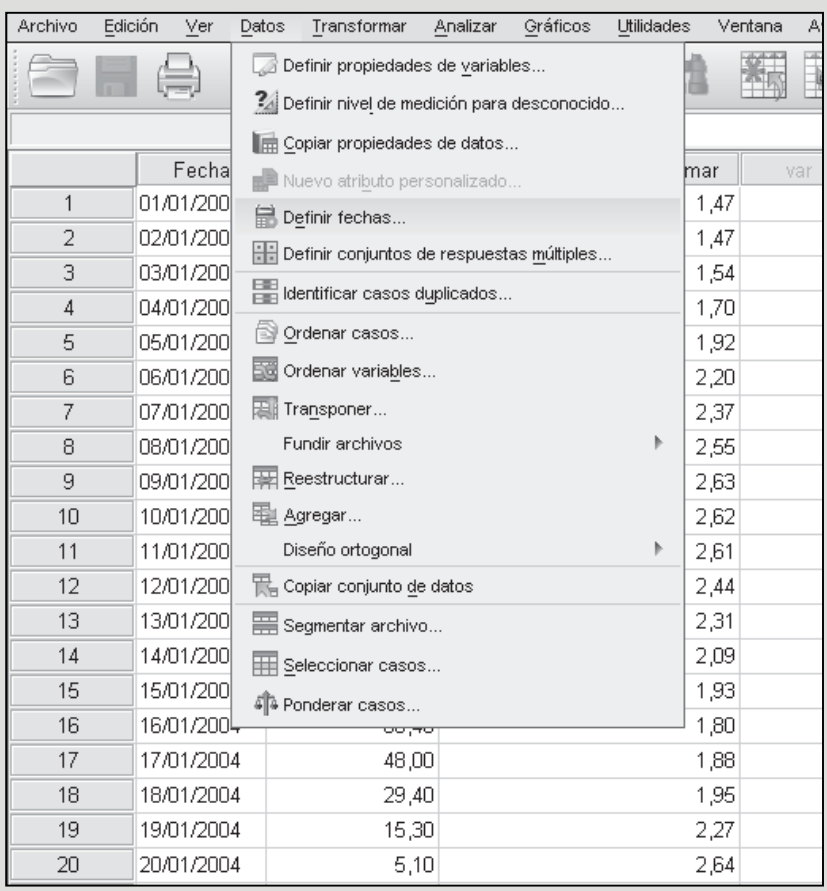
Como la serie de datos es desde 1925 a 1996, solo 71 años, no es posible determinar si el otro ciclo que se había observado, el ciclo de Gleissberg, que tenía un periodo promedio de 86,3 años, también influye en la temperatura del océano.

Cuadro XII.2. *(Continuación)*[SPSS]

Aplicación con SPSS.

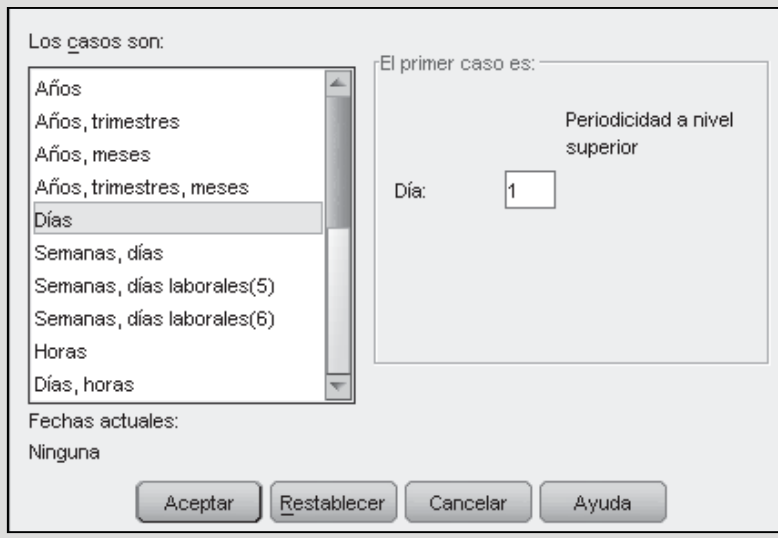
EJEMPLO. Vamos a utilizar los mismos datos que en la aplicación con R. En el archivo **Cuadro.XII.2.sav** están los ángulos entre el sol y la luna y la diferencia en metros entre pleamar y bajamar para 42º 8´Norte y 14º 13´Oeste del 1/1/2004 al 14/3/2004. El primer objetivo es determinar el ciclo de los cambios en el ángulo entre el sol y la luna. El segundo objetivo es mostrar la influencia de la posición relativa del sol y la luna en las diferencias entre pleamar y bajamar.

Paso 1. Aunque ya existan unas variables informando sobre el tiempo, en este ejemplo «Fecha», hay que definir unas nuevas para poder realizar el análisis. Para ello vamos a «Datos» ⇨ «Definir fechas...».

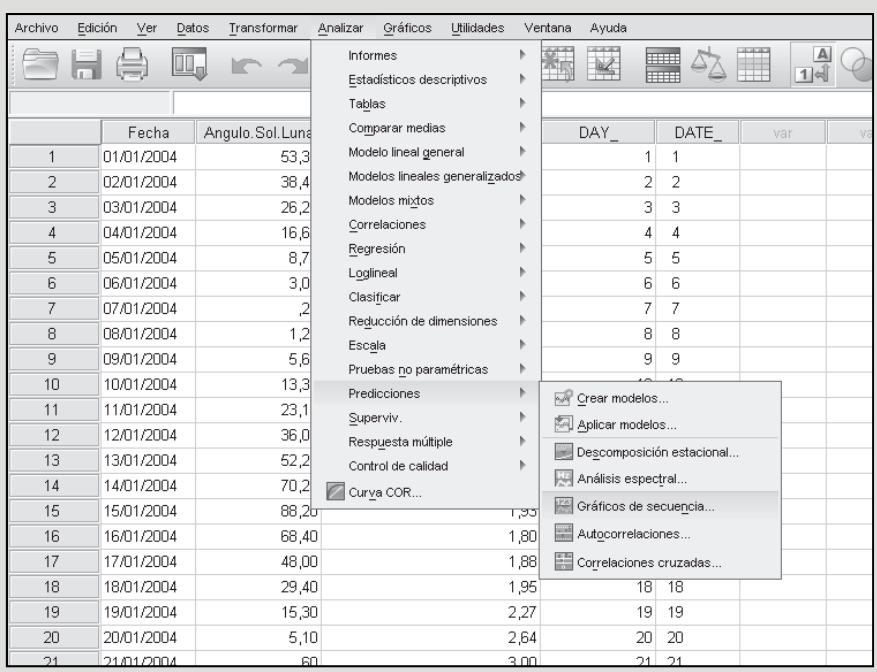


Cuadro XII.2. *(Continuación)*[SPSS]

Paso 2. Aparece la siguiente ventana donde elegimos «Días» ya que nuestra serie temporal está expresada en esas unidades de tiempo.

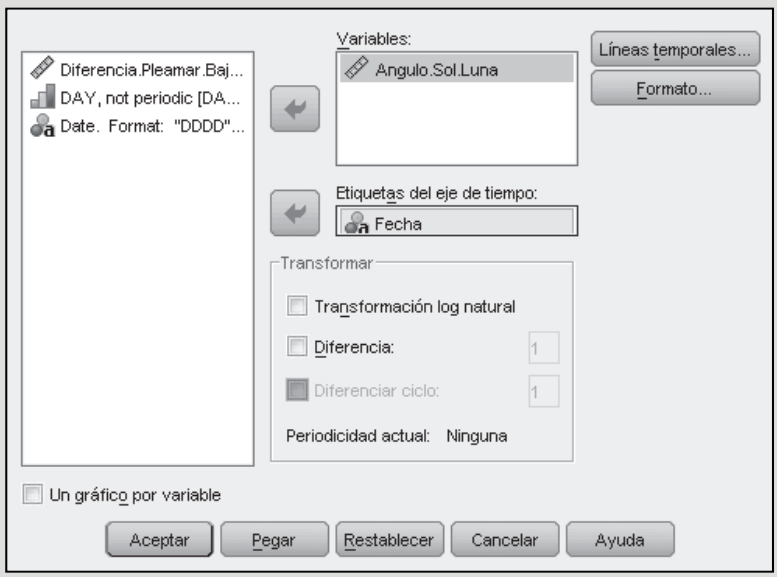


Paso 3. A continuación representamos la variable «Angulo.Sol.Luna» para tener una idea de como varía en el tiempo. Para ello seleccionamos «Analizar» ⇨ «Predicciones» ⇨ «Gráficos de secuencia...».

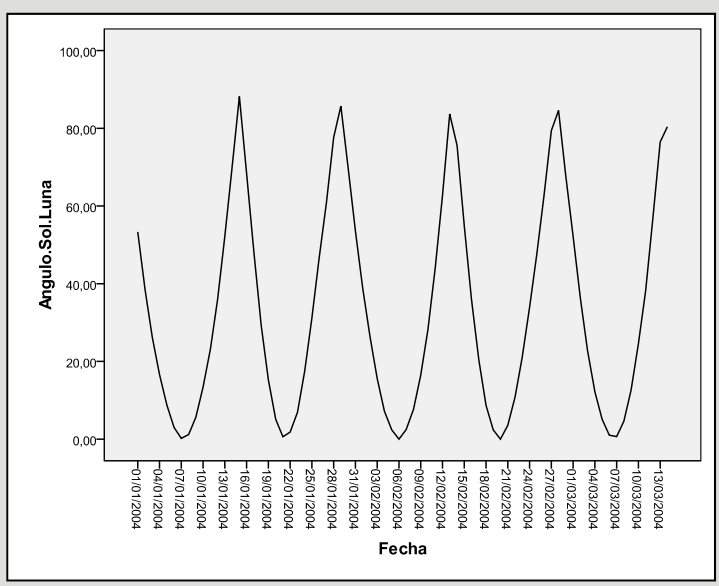


Cuadro XII.2. *(Continuación)*[SPSS]

Paso 4. Aparece la siguiente ventana donde se especifica la variable a representar, en este caso «Angulo.Sol.Luna», frente al tiempo expresado en días.

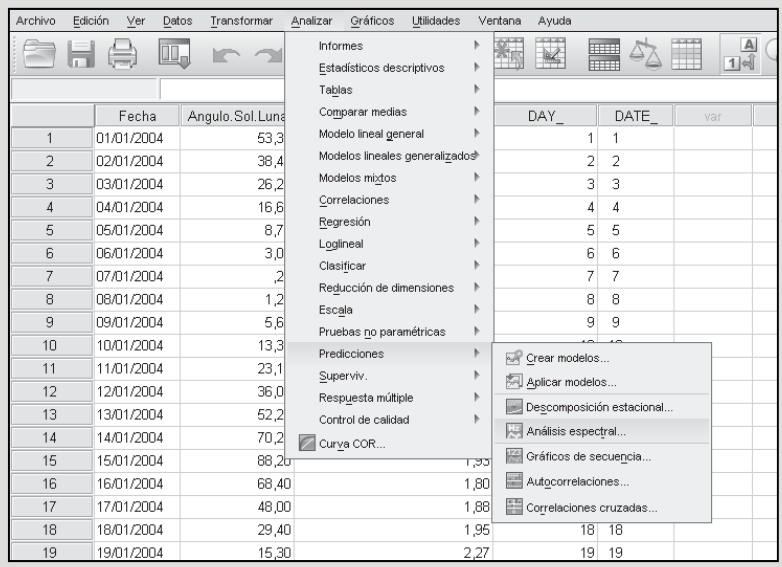


Paso 5. La gráfica que se obtiene muestra una clara variación cíclica, con 5 oscilaciones en el conjunto de los 2,5 meses, lo que corresponde a un período de aproximadamente 15 días. Para profundizar un poco más realizaremos un análisis espectral.

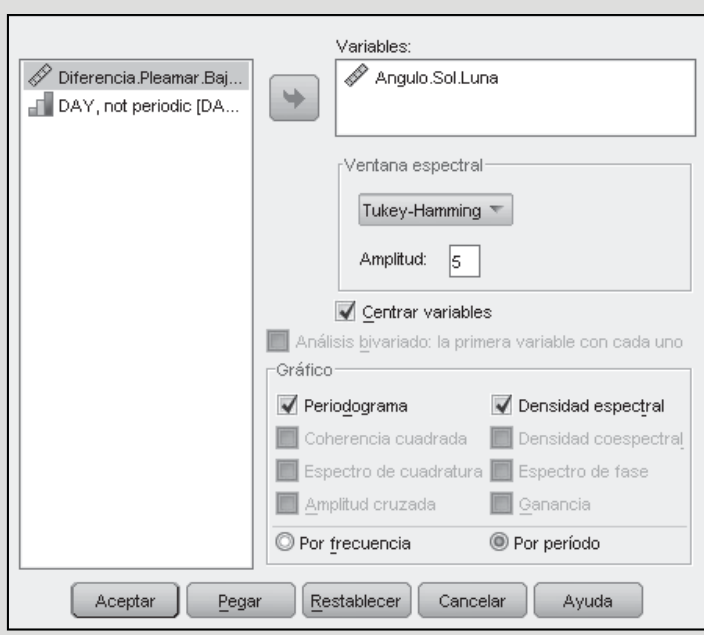


Cuadro XII.2. *(Continuación)*[SPSS]

Paso 6. Se selecciona «Analizar» ⇨ «Predicciones» ⇨ «Análisis espectral...».



Paso 7. Introduciremos la variable de interés «Angulo.Sol.Luna» y elegimos los gráficos «Periodograma» y «Gráfico de densidad espectral». Por último, en «Gráfico» se selecciona «Por período».



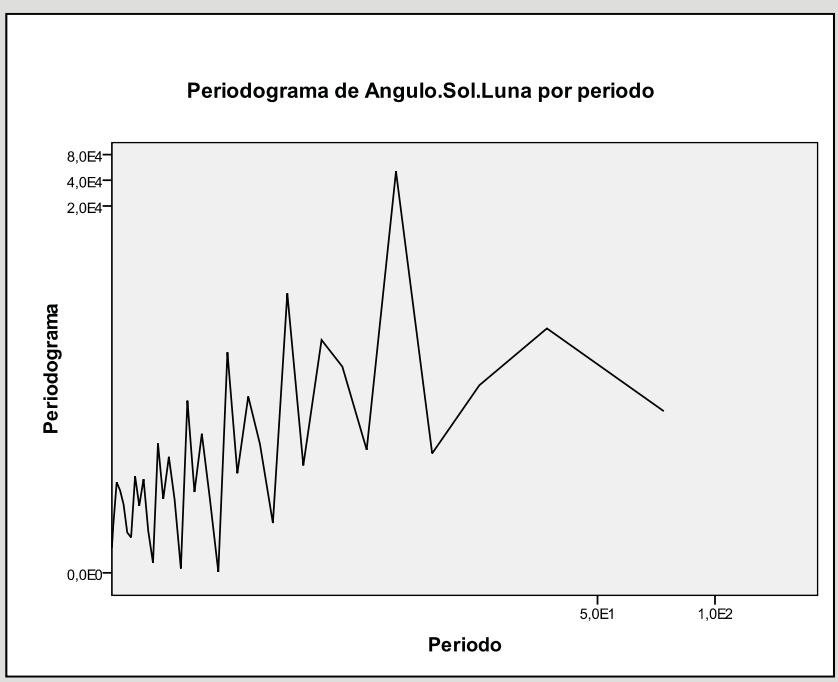
Cuadro XII.2. *(Continuación)*[SPSS]

El gráfico de densidad espectral es simplemente una versión suavizada del periodograma, eliminando mediante filtrado el ruido formado por las variaciones aleatorias, o en general cualquier variación de alta frecuencia. Generalmente construiremos ambos gráficos y utilizaremos el que presente los resultados más claros.

En el apartado «Ventana espectral» aparecen varios procedimientos para realizar el suavizado o filtrado y obtener la función de densidad espectral (cualquiera de ellos es válido, por lo que podemos mantener el de Tukey-Hamming que aparece por defecto).

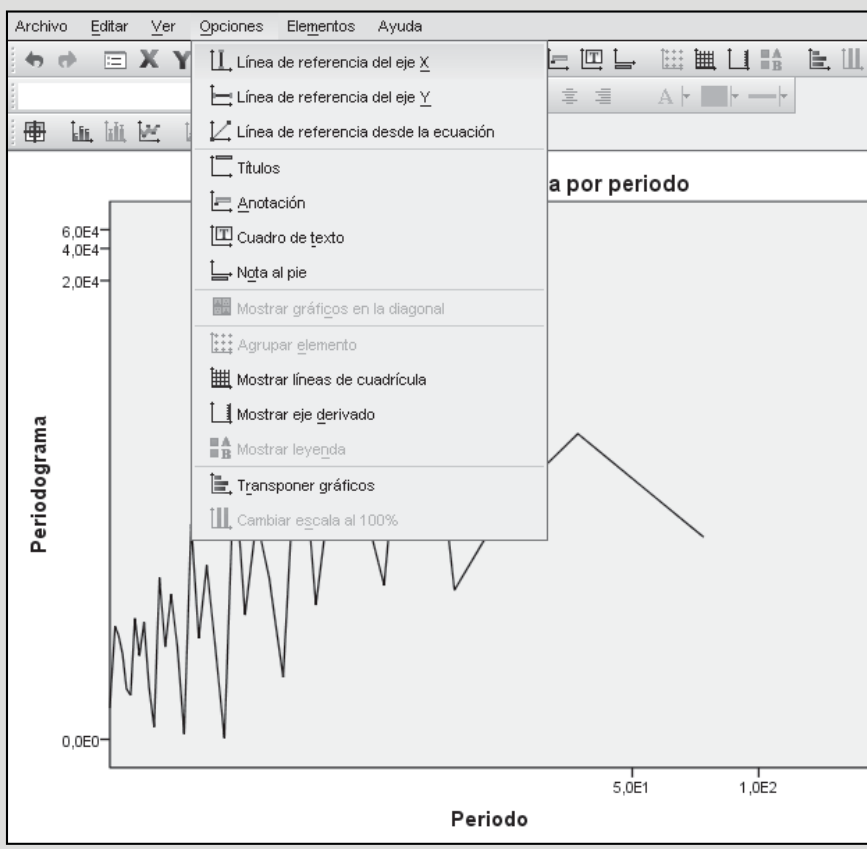
Ambos gráficos pueden hacerse «Por frecuencia» o «Por período». En el primer caso aparecen en el eje horizontal todas las frecuencias, expresadas como fracción del espacio de tiempo total de la serie (en una serie de 20 años una fracción de 0,05 equivale a 1 año), y en el segundo se mostrarán los distintos períodos, expresados en la unidad de tiempo de referencia de los datos (en nuestro ejemplo en días). Elegimos esta segunda opción –por período–, que nos permite interpretar los resultados directamente.

Paso 8. En el gráfico resultante observamos un pico que destaca claramente sobre los demás. Esto indica la existencia de un único ciclo, ya que si hubiese varios superpuestos habría distintos picos o máximos reconocibles claramente en el gráfico. El paso siguiente consiste en la identificación de la longitud o período de ese ciclo.



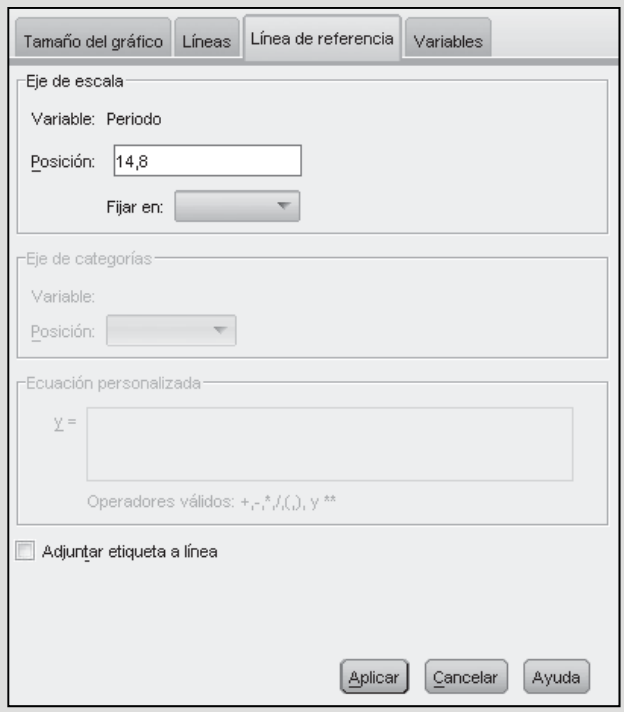
Cuadro XII.2. *(Continuación)*[SPSS]

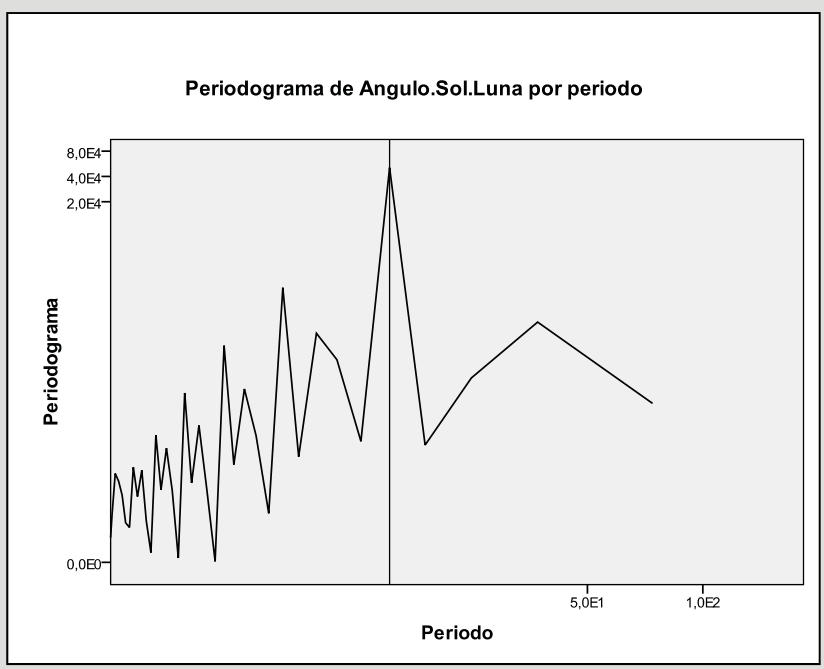
Paso 9. La escala horizontal del gráfico es muy limitada (aparecen indicados solamente los valores 50 y 100 días) y además es logarítmica, lo que no facilita su lectura directa. Una escala natural, no logarítmica, en el eje de tiempo sería todavía más confusa en series largas, pero tenemos algunas posibilidades de interpretación editando el gráfico para saber cual es el período del ciclo que hemos identificado. Para ello, editamos el gráfico con un doble clic, con lo que se abre el editor de gráficos y se selecciona «Opciones» ⇨ «Añadir una línea de referencia al eje X» para incluir referencias adicionales.



Paso 10. Como el tiempo está medido en días y hemos identificado en el gráfico de secuencia un período aproximado de 15 días, indicaremos en «Posición» del eje diferentes valores alrededor de ese ciclo y pulsamos «Aplicar». Así se va probando hasta obtener un periodo de tiempo que coincida aproximadamente con el máximo del gráfico. Se observa que el ciclo tiene un período o duración media de unos 14,8 días.

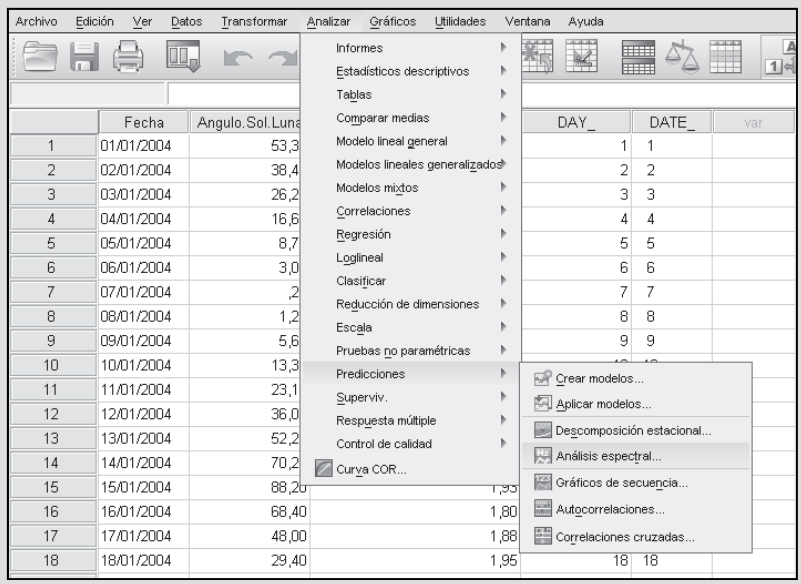
Cuadro XII.2. *(Continuación)*[SPSS]



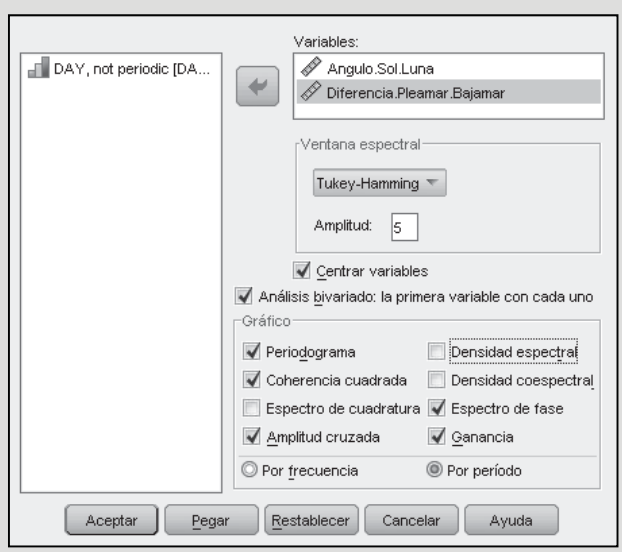


Cuadro XII.2. *(Continuación)*[SPSS]

Paso 11. Para el segundo objetivo, es decir, determinar la posible existencia de una relación entre el ángulo que forman el sol y la luna con la intensidad de las mareas, se selecciona «Analizar» ⇨ «Predicciones» ⇨ «Análisis espectral...».



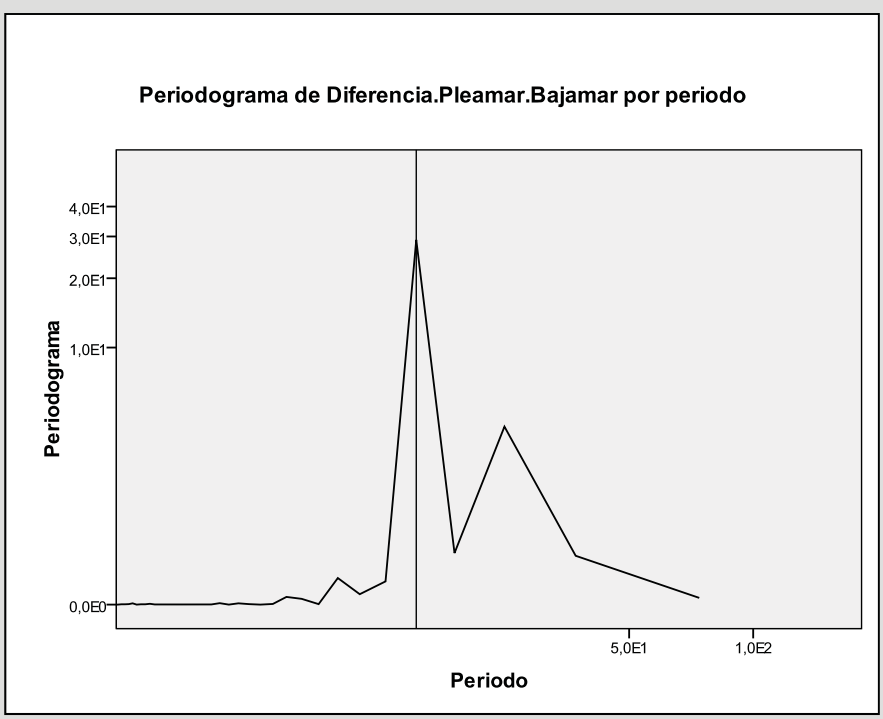
Paso 12. En el cuadro de diálogo que aparece introducimos las variables objetivo. La primera debe ser la variable independiente o explicativa, y la segunda la variable dependiente. En nuestro ejemplo, la hipótesis es que la diferencia en la altura de las mareas depende de la posición relativa del sol y la luna, por lo que ponemos «Angulo.Sol.Luna» en primer lugar.



Cuadro XII.2. *(Continuación)*[SPSS]

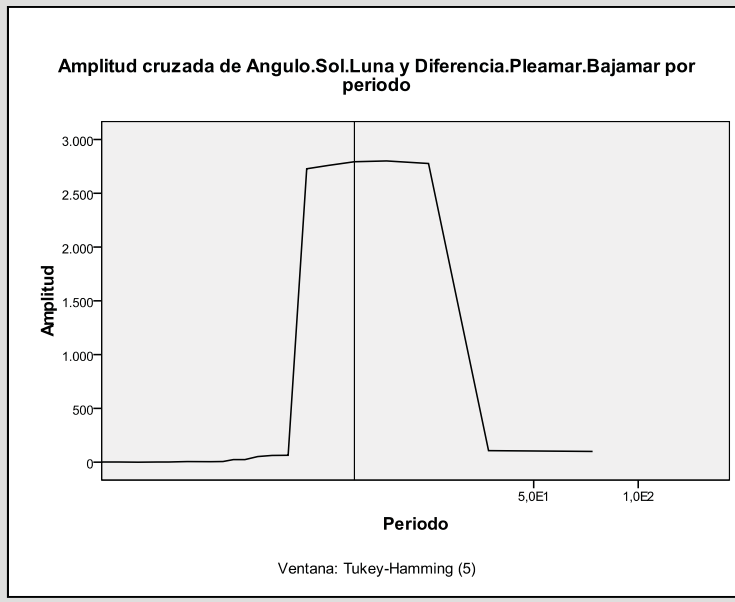
En la ventana anterior se marca la casilla «Análisis bivariados: la primera variable con cada uno», con lo cual se activan las opciones de gráficos. Marcamos «Por período», más fácil de interpretar que «Por frecuencia», y elegimos los estadísticos de interés: «Periodograma», «Coherencia cuadrada», «Amplitud cruzada», «Espectro de fase» y «Ganancia». Pulsamos «Aceptar» para obtener los resultados.

Paso 13. Anteriormente ya vimos que el ciclo de la posición del sol y la luna tenía una periodicidad de aproximadamente 14,8 días. Ahora analizamos en segundo lugar el periodograma de la variable dependiente «Diferencia.Pleamar.Bajamar». Para ello hacemos lo mismo que se explicó en los pasos 9 y 10. Se observa claramente un pico predominante en la misma duración (14,8 días), coincidente con la variable anterior. Se trata, por lo tanto, de una variable cíclica. Parece existir otro ciclo de mayor duración, situado a la derecha del anterior, pero mucho menos importante, por lo que no lo tendremos en cuenta en nuestro análisis.

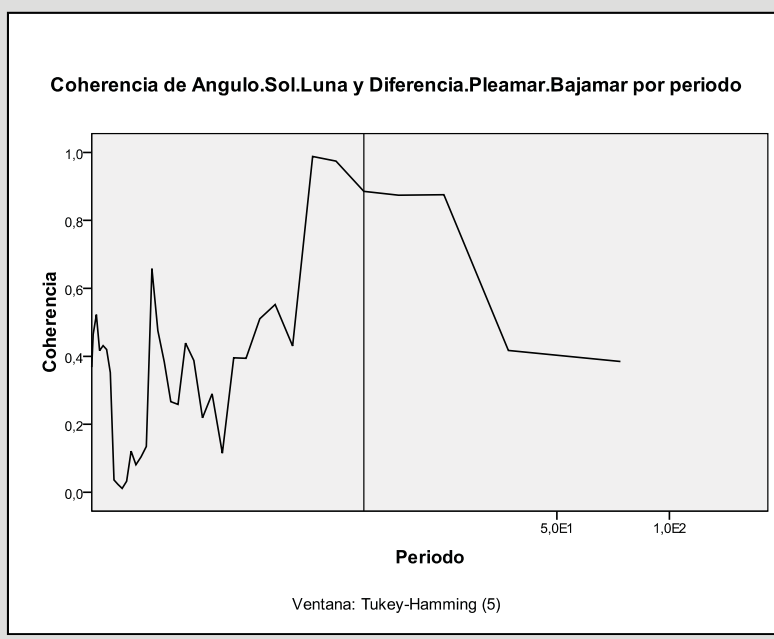


Paso 14. La amplitud cruzada, que se muestra en la gráfica siguiente, presenta los valores más altos para el mismo período, evidenciando la evolución sincronizada de ambas variables.

Cuadro XII.2. *(Continuación)*[SPSS]

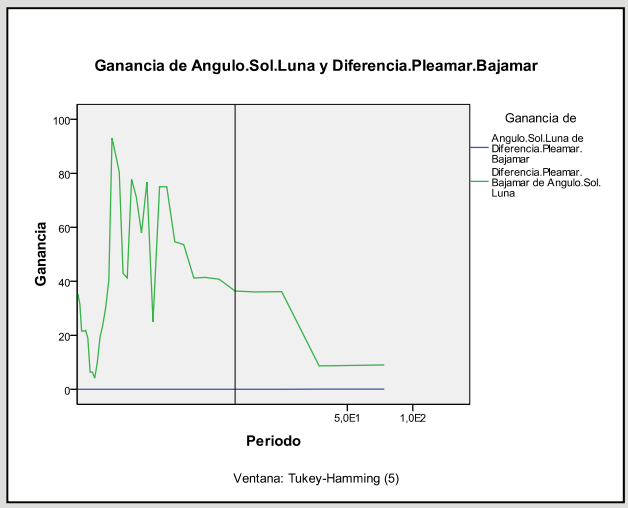


Paso 15. La coherencia mide el cuadrado de la correlación entre ambas series para cada frecuencia. Un valor próximo a 1, como el que presentan nuestras dos variables para el período de 14,8 días, indica una clara asociación entre las mareas y la posición relativa del sol y la luna. Es el estadístico de mayor interés en la interpretación de la relación entre dos variables cíclicas.

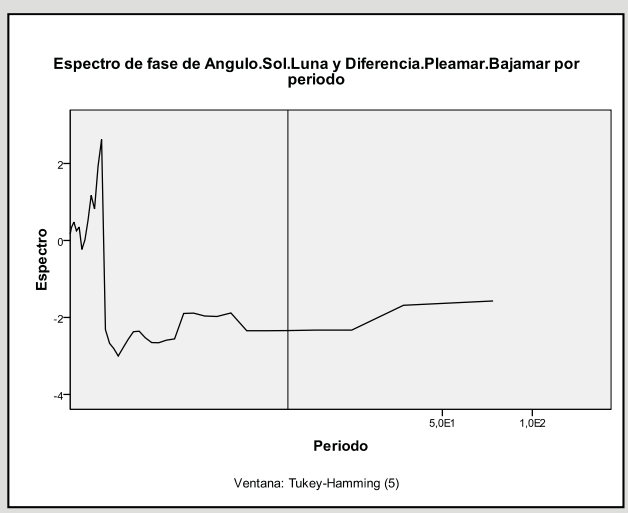


Cuadro XII.2. *(Continuación)*[SPSS]

Paso 16. La ganancia es otro indicador de interés, ya que es el número por el que se multiplica la amplitud de una variable para reproducir la componente de amplitud de la otra variable para cada frecuencia. Este estadístico es como la pendiente en una regresión lineal: cuanto más grande, más relación hay. En nuestro ejemplo observamos valores relativamente grandes, en torno a 40, para el período de 14,8 días.



Paso 17. Por último, una vez constatada la relación entre los ciclos de ambas variables, interesa averiguar si existe algún retardo o desfase en la asociación. Para ello se construye el espectro de fase, con el retardo entre una serie y otra para cada frecuencia. Observamos en el siguiente gráfico como para el período ya conocido de 14,8 días existe un desfase de aproximadamente -2 días. Por tanto, el efecto sobre las mareas se produce con casi dos días de retraso sobre los cambios en la posición relativa del sol y la luna.



Capítulo XIII

ANÁLISIS MULTIVARIANTE: MÉTODOS FACTORIALES

El análisis multivariante consiste en un conjunto de técnicas estadísticas cuya finalidad es el análisis de datos en los que contamos con más de dos variables medidas para cada caso estudiado.

En estas situaciones, a menudo con muchas variables, la interpretación con el análisis uni y bidimensional resulta muy limitada, ya que impide considerar toda la información simultáneamente, e ignora los efectos conjuntos o interacciones entre varias variables. En el estudio de la realidad nuestros datos son casi siempre multivariantes, y la mayor complejidad matemática de los modelos o las dificultades de cálculo –resueltas por medio de computadoras– no debe impedir la utilización de las técnicas más adecuadas y potentes.

Los modelos con varias variables en los que una sola de ellas se considera dependiente de las demás (como regresión múltiple), se tratan a menudo dentro del conjunto de técnicas univariantes. En este capítulo nos centraremos en las técnicas multivariantes más útiles y frecuentes, que podemos agrupar en dos grandes categorías: métodos factoriales y métodos de clasificación (estos últimos serán tratados en el capítulo siguiente).

Los métodos factoriales, por extensión del Análisis Factorial, se caracterizan por estar basados en el modelo lineal -el modelo matemático más simple-, su resolución utiliza técnicas algebraicas, y la mayoría de ellos tienen como principal resultado un gráfico que permite interpretar de forma sencilla un conjunto complejo de datos. Veremos dentro de este grupo el Análisis de Componentes Principales, Análisis Factorial, Análisis de Correspondencias y Escalamiento multidimensional.

XIII.1. Análisis de Componentes Principales

Consiste en la transformación del conjunto de variables originales en otro conjunto de variables –Componentes Principales– obtenidas como combinación lineal de aquellas. Las nuevas variables, en igual número que las antiguas, conservan toda la información –en lo que se refiere a la variabilidad del conjunto de objetos– de las variables primitivas, pero la mayoría de las componentes principales tienen una variabilidad tan pequeña que pueden ser ignoradas, de tal manera que unas pocas componentes –generalmente tres o menos– permiten representar y explicar razonablemente el conjunto de objetos de la muestra sin pérdida importante de información.

La principal virtud del Análisis de Componentes Principales (ACP) consiste precisamente en la reducción de la complejidad de los datos, al pasar de muchas variables a pocas (frecuentemente una o dos), que pueden ser representadas gráficamente. Frente a esta ventaja, el inconveniente de las nuevas variables está en la pérdida parcial de información (a menudo pequeña) y en el hecho de

que las variables originales tienen un significado real del que carecen en general las componentes, que son mezcla o combinación de variables.

La primera componente principal es la combinación lineal de las variables originales que tiene máxima varianza. La segunda componente es la combinación lineal de las variables originales con máxima varianza con la condición añadida de que sea independiente de la primera (ortogonal), y así sucesivamente se pueden obtener todas las componentes principales, que siendo independientes entre sí contienen información distinta, ya que la independencia o ausencia de correlación significa que las nuevas variables o componentes no comparten información común. Cada componente principal explica por lo tanto la máxima variabilidad residual posible (la que no han explicado ya las anteriores).

Para encontrar las combinaciones óptimas de las variables originales deben ser calculados los autovalores y autovectores de la matriz de covarianzas (o de la matriz de correlaciones). Los autovalores, ordenados de mayor a menor, son las varianzas de las nuevas componentes, y los autovectores expresan las combinaciones lineales que definen las componentes principales.

Debemos decidir cuantas componentes deben ser conservadas: un número alto permite explicar una mayor proporción de la variabilidad total, y un número bajo permite obtener una mayor simplicidad en la representación gráfica. Existen varios criterios que pueden ayudarnos a la hora de pronunciarnos:

- *Criterio de la varianza*. Se seleccionan las componentes que explican un porcentaje suficiente (por ejemplo 80% o 90%) de la varianza.
- *Criterio de Kaiser*. Se conservan las componentes cuyas varianzas (autovalores) son mayores que la varianza media.
- *Gráfico de sedimentación* (*Scree plot)*. Representamos los autovalores, que en general decrecen rápidamente al principio y lentamente después, y conservamos únicamente los que se sitúan por encima del codo de la curva.
- *Dos componentes.* Mantener solamente las dos primeras permite construir un gráfico en dos dimensiones, generalmente el más adecuado para describir el conjunto de datos.

La única condición para que pueda aplicarse el Análisis de Componentes Principales es que las variables sean cuantitativas. Sin embargo, existe otra condición o hipótesis necesaria para que sea útil su aplicación: las variables originales deben estar correlacionadas. Cuanto más alta sea la correlación entre las variables, mayor es la proporción de varianza explicada por las primeras componentes y, por lo tanto, más pequeña será la pérdida de información al conservarlas prescindiendo de las restantes.

Para comprobar si nuestras variables están correlacionadas existen varios procedimientos:

- *Inspeccionar la matriz de correlaciones*. Podremos observar si la mayoría de los coeficientes de correlación lineal entre las variables originales son próximos a 1 o a -1.
- *Inspeccionar la matriz de valores p del contraste de nulidad de los coeficientes de correlación*. Se contrasta –para cada uno de los coeficientes de correlación– la hipótesis de que es cero. El valor *p* debe ser menor que el nivel de significación (habitualmente 0,05) para rechazar esa hipótesis. De este modo, la matriz de valores *p* cuando existe correlación debe estar formada mayoritariamente por ceros o valores muy pequeños.

- *Determinante de la matriz de correlaciones*. Toma un valor próximo a 1 cuando no existe correlación y próximo a cero cuando existe.
- *Test de esfericidad de Bartlett*. Está basado en el determinante de la matriz de correlaciones y contrasta la hipótesis de "esfericidad" o ausencia de correlación con un único test. Un valor *p* del contraste menor que el nivel de significación elegido permite rechazar la hipótesis y concluir que existe correlación.

Cuando las variables originales son heterogéneas, es decir, vienen expresadas en unidades de medida diferentes, ocurre que la variabilidad viene determinada por la elección de la unidad de medida de cada una de ellas, lo que influye en los resultados de la aplicación del ACP al asignar una importancia arbitraria a cada variable. Para evitar esta influencia negativa, la solución consiste en utilizar la matriz de correlaciones en lugar de la matriz de covarianzas, lo que equivale a tipificar las variables originales. De este modo todas las variables pasan a tener media 0 y desviación típica 1. Como este proceso implica una cierta pérdida de información, no debemos utilizar la matriz de correlaciones en sustitución de la matriz de covarianzas en el caso de que nuestras variables sean homogéneas.

Una vez obtenidas las combinaciones lineales que definen a las componentes principales, podemos interpretar los coeficientes (saturaciones o cargas): cuanto más alto sea su valor absoluto, mayor influencia tiene esa variable en la construcción de la componente. Es posible interpretar las componentes, y asignarles significados genéricos, a partir del conjunto de variables con mayor peso en su construcción.

Para facilitar esta interpretación de las componentes y atribuirles algún significado puede ser útil el gráfico de saturaciones, que muestra –en el plano de las dos primeras componentes– los coeficientes o saturaciones para cada variable, de modo que cada variable original se representa como un punto en ese plano. Las variables que están más alejadas con respecto al cero del eje horizontal son las que dan significado a la primera componente (que a menudo expresa una idea genérica de "tamaño"), y las que están más alejadas respecto al cero del eje vertical son las que dan sentido a la segunda componente (a menudo refleja una idea genérica de "forma"). Las variables próximas al centro de coordenadas no tienen interés, y las que se alejan de ambos ejes son comunes a ambas componentes y tampoco ayudan a su interpretación. El gráfico puede hacerse también para otras componentes distintas de las dos primeras.

El principal resultado de esta técnica es en general el gráfico de puntuaciones. Consiste en la representación, en el plano de las dos primeras componentes, de las coordenadas de los puntos que representan a los elementos del conjunto objeto (las filas o casos en nuestros datos). Esta representación permite describir el conjunto de datos multidimensionales de un modo relativamente sencillo. Puede ser útil añadir al gráfico etiquetas de caso o marcas.

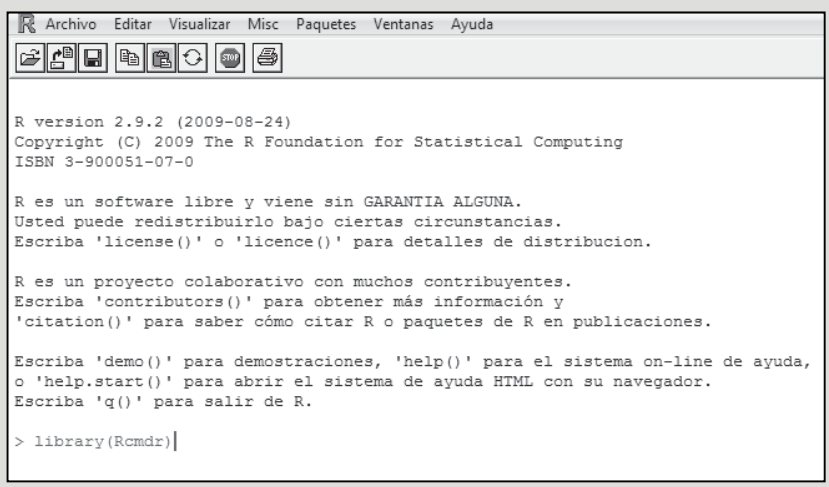
En el Cuadro XIII.1 se dan varios ejemplos de ACP usando R, STATISTICA y SPSS.

CUADRO XIII.1. Componentes principales

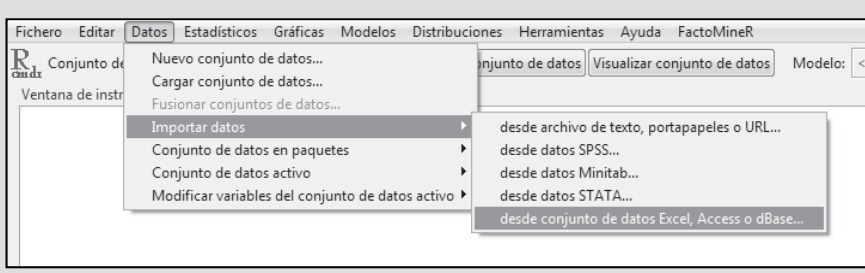
Aplicación con R.

EJEMPLO. Tenemos datos de un conjunto de clientes de un banco, los cuales están en el archivo **Cuadro XIII.1.xls**. Las variables indican el nivel de ingresos, la relación (en porcentaje) entre deudas e ingresos, deudas con tarjeta de crédito, y otras deudas. Además existen otras variables cualitativas, que establecen grupos de edad, nivel de estudios, antigüedad en el empleo, antigüedad en la residencia actual, y si se han producido o no impagos. Las variables cuantitativas indican la posición financiera de cada cliente y queremos asociar esa posición con las diferentes variables cualitativas.

Paso 1. Iniciamos R y cargamos el paquete «Rcmdr» introduciendo la orden «library(Rcmdr)». Si el paquete no está instalado, la instalación se realiza como se explica en el Apéndice II en el apartado AII.IV.

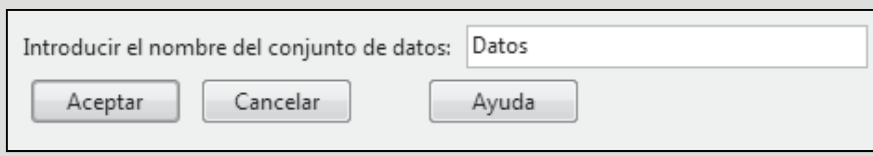

Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> library(Rcmdr)

Paso 2. Después de haber seleccionado el directorio de trabajo, como se explica en el Apéndice II, vamos a «Datos» ⇨ «Importar datos» ⇨ «desde conjunto de datos Excel, Access o dBase».

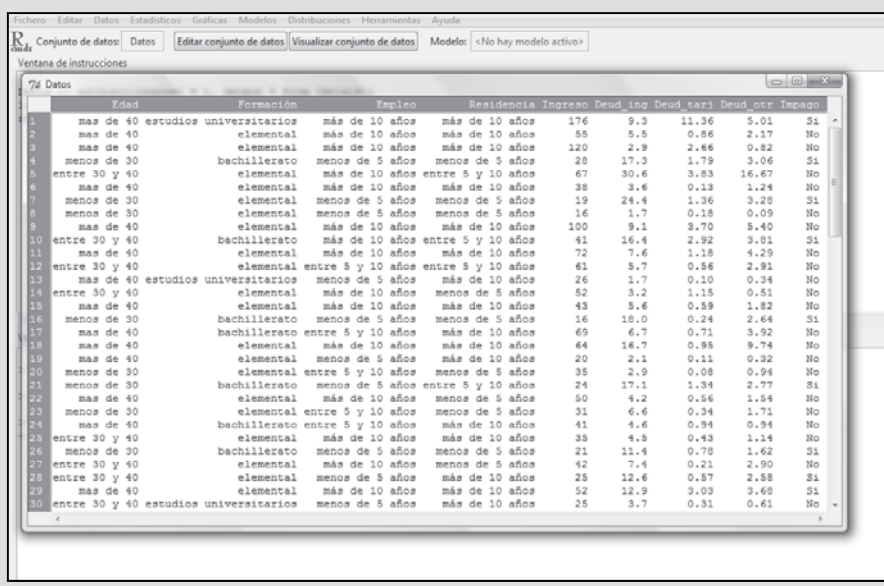

Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda FactoMineR
Nuevo conjunto de datos...
Cargar conjunto de datos...
Fusionar conjuntos de datos...
Importar datos
Conjunto de datos en paquetes
Conjunto de datos activo
Modificar variables del conjunto de datos activo
desde archivo de texto, portapapeles o URL...
desde datos SPSS...
desde datos Minitab...
desde datos STATA...
desde conjunto de datos Excel, Access o dBase...
Visualizar conjunto de datos
Modelo:

CUADRO XIII.1. *(Continuación)*[R]

Paso 3. Aparece un pequeño cuadro de diálogo para introducir el nombre del conjunto de datos, en el que dejamos el nombre por defecto «Datos» y pulsamos «Aceptar». Posteriormente aparecen los archivos del directorio de trabajo, en el que elegimos **Cuadro XIII.1.xls** y pulsamos «Abrir».



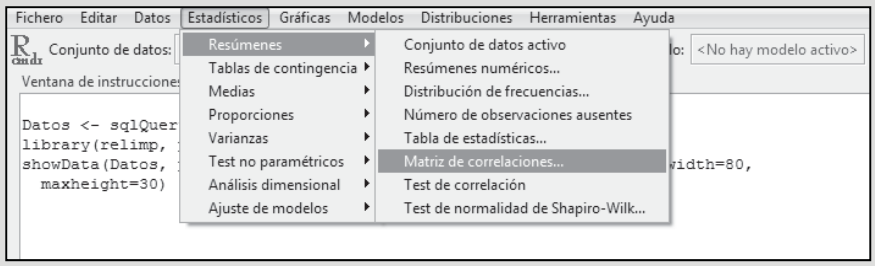
Paso 4. En R Commander pulsamos en «Visualizar conjunto de datos». Es importante inspeccionar los datos y comprobar que no existan líneas, lo cual a veces ocurre al final del archivo, que tengan celdas con el código <NA>. Si existen estas líneas, hay que eliminarlas.



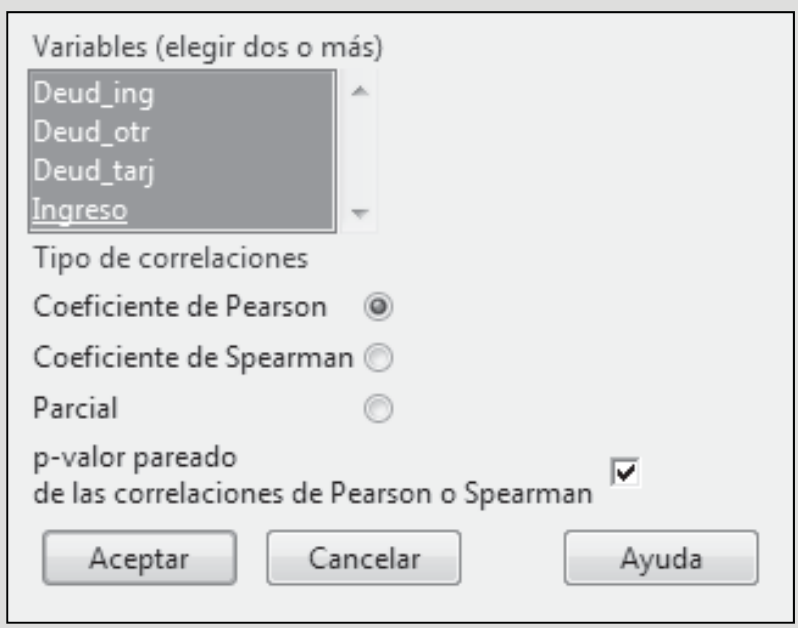
	Edad	Formación	Empleo	Residencia	Ingreso	Deud_ing	Deud_tarj	Deud_otr	Impago
1	mas de 40	estudios universitarios	más de 10 años	más de 10 años	176	9.3	11.36	5.01	Si
2	mas de 40	elemental	más de 10 años	más de 10 años	55	5.5	0.86	2.17	No
3	mas de 40	elemental	más de 10 años	más de 10 años	120	2.9	2.66	0.82	No
4	menos de 30	bachillerato	menos de 5 años	menos de 5 años	28	17.3	1.79	3.06	Si
5	entre 30 y 40	elemental	más de 10 años	entre 5 y 10 años	67	30.6	3.83	16.67	No
6	mas de 40	elemental	más de 10 años	más de 10 años	38	3.6	0.13	1.24	No
7	menos de 30	elemental	menos de 5 años	menos de 5 años	19	24.4	1.36	3.28	Si
8	menos de 30	elemental	menos de 5 años	menos de 5 años	16	1.7	0.18	0.09	No
9	mas de 40	elemental	más de 10 años	más de 10 años	100	9.1	3.70	5.40	No
10	entre 30 y 40	bachillerato	más de 10 años	entre 5 y 10 años	41	16.4	2.92	3.81	Si
11	mas de 40	elemental	más de 10 años	más de 10 años	72	7.6	1.18	4.29	No
12	entre 30 y 40	elemental	entre 5 y 10 años	entre 5 y 10 años	61	5.7	0.56	2.91	No
13	mas de 40	estudios universitarios	menos de 5 años	más de 10 años	26	1.7	0.10	0.34	No
14	entre 30 y 40	elemental	más de 10 años	menos de 5 años	52	3.2	1.15	0.51	No
15	mas de 40	elemental	más de 10 años	más de 10 años	43	5.6	0.59	1.82	No
16	menos de 30	bachillerato	menos de 5 años	menos de 5 años	16	18.0	0.24	2.64	Si
17	mas de 40	bachillerato	entre 5 y 10 años	más de 10 años	69	6.7	0.71	3.92	No
18	mas de 40	elemental	más de 10 años	más de 10 años	64	16.7	0.95	9.74	No
19	mas de 40	elemental	menos de 5 años	más de 10 años	20	2.1	0.11	0.32	No
20	menos de 30	elemental	entre 5 y 10 años	menos de 5 años	35	2.9	0.08	0.94	No
21	menos de 30	bachillerato	menos de 5 años	entre 5 y 10 años	24	17.1	1.34	2.77	Si
22	mas de 40	elemental	más de 10 años	menos de 5 años	50	4.2	0.56	1.54	No
23	menos de 30	elemental	entre 5 y 10 años	menos de 5 años	31	6.6	0.34	1.71	No
24	mas de 40	bachillerato	entre 5 y 10 años	más de 10 años	41	4.6	0.94	0.94	No
25	entre 30 y 40	elemental	más de 10 años	más de 10 años	38	4.5	0.43	1.14	No
26	menos de 30	bachillerato	menos de 5 años	menos de 5 años	21	11.4	0.78	1.62	Si
27	entre 30 y 40	elemental	más de 10 años	menos de 5 años	42	7.4	0.21	2.90	No
28	entre 30 y 40	elemental	menos de 5 años	más de 10 años	25	12.6	0.57	2.58	Si
29	mas de 40	elemental	más de 10 años	más de 10 años	52	12.9	3.03	3.68	Si
30	entre 30 y 40	estudios universitarios	menos de 5 años	más de 10 años	25	3.7	0.31	0.61	No

Paso 5. Antes de aplicar el Análisis de Componentes Principales debemos comprobar si nuestras variables están correlacionadas. Pulsamos en el menú principal «Estadísticos» ⇨ «Resúmenes» ⇨ «Matriz de correlaciones».

CUADRO XIII.1. *(Continuación)*[R]



Paso 6. En el cuadro de diálogo que aparece seleccionamos todas las variables, y marcamos «Coeficiente de Pearson» y «p-valor pareado», y pulsamos «Aceptar». Solo aparecen las variables cuantitativas.



Paso 7. En la ventana de resultados de R Commander aparece la matriz de correlaciones, la de valores *p* de los contrastes de nulidad, y otra de valores *p* corregidos para múltiples contrastes. Es más correcto utilizar esta última cuando exploramos la existencia de posibles correlaciones nulas, y la primera cuando solo nos interesa contrastar un único coeficiente de correlación concreto.

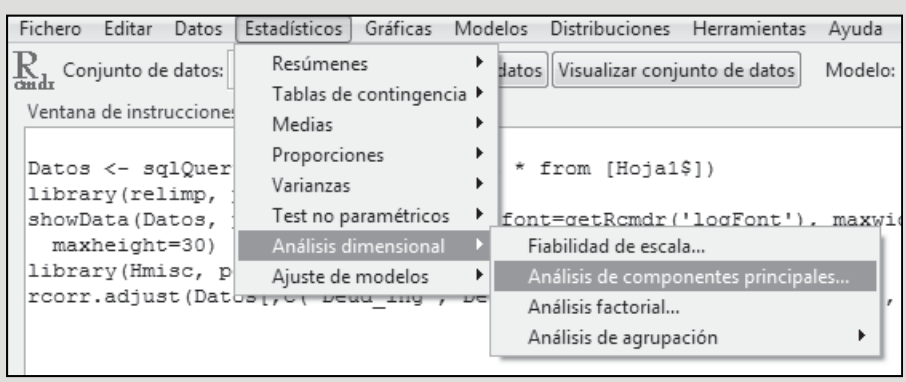
Como podemos observar en la matriz de valores *p* corregidos «Adjusted p-values», todos son menores que 0,05 lo que indica que ninguno de los coeficientes de correlación parece ser nulo. Podemos observar también los valores de los coeficientes de correlación, los cuales no son muy elevados a pesar de ser significativas las correlaciones (0,65, 0,44, 0,5, etc.).

CUADRO XIII.1. *(Continuación)*[R]

Ventana de resultados

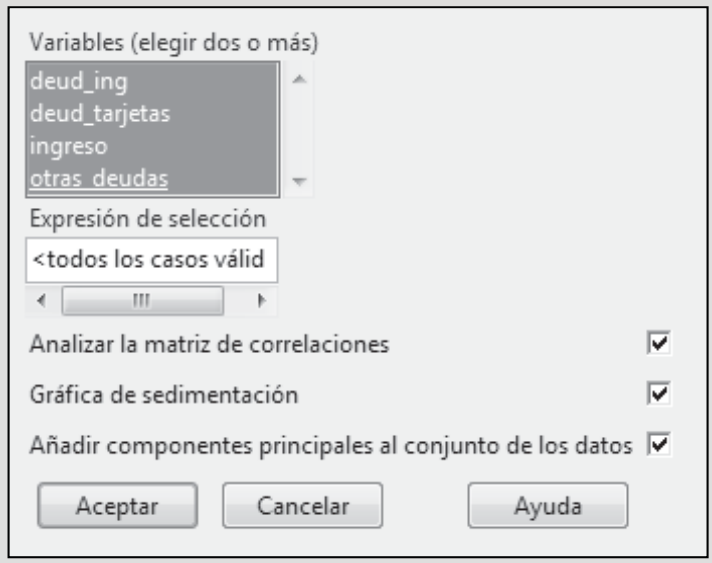
```
+   maxheight=30)

> library(Hmisc, pos=4)

> rcorr.adjust(Datos[,c("Deud_ing","Deud_otr","Deud_tarj","Ingreso")], type="pearson")
          Deud_ing Deud_otr Deud_tarj Ingreso
Deud_ing      1.00     0.65      0.44   -0.20
Deud_otr      0.65     1.00      0.55    0.33
Deud_tarj     0.44     0.55      1.00    0.54
Ingreso      -0.20     0.33      0.54    1.00

n= 100


P
          Deud_ing Deud_otr Deud_tarj Ingreso
Deud_ing           0.0000   0.0000    0.0414
Deud_otr  0.0000            0.0000    0.0008
Deud_tarj 0.0000   0.0000             0.0000
Ingreso   0.0414   0.0008   0.0000

 Adjusted p-values (Holm's method)
          Deud_ing Deud_otr Deud_tarj Ingreso
Deud_ing           0.0000   0.0000    0.0414
Deud_otr  0.0000            0.0000    0.0016
Deud_tarj 0.0000   0.0000             0.0000
Ingreso   0.0414   0.0016   0.0000
```

Paso 8. Para aplicar el Análisis de Componentes Principales elegimos en el menú principal de R Commander las opciones «Estadísticos» ⇨ «Análisis dimensional» ⇨ «Análisis de componentes principales».

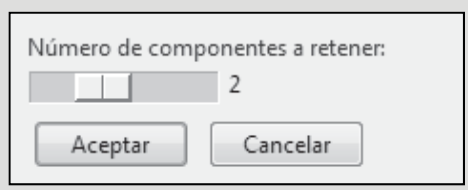


Paso 9. En el cuadro de diálogo que aparece seleccionamos las variables que queremos utilizar (en este caso todas), marcamos «Analizar la matriz de correlaciones», dado que algunas variables están expresadas en euros y otras en porcentaje (son heterogéneas). Se marca también «Gráfica de sedimentación» y «Añadir componentes principales al conjunto de datos».

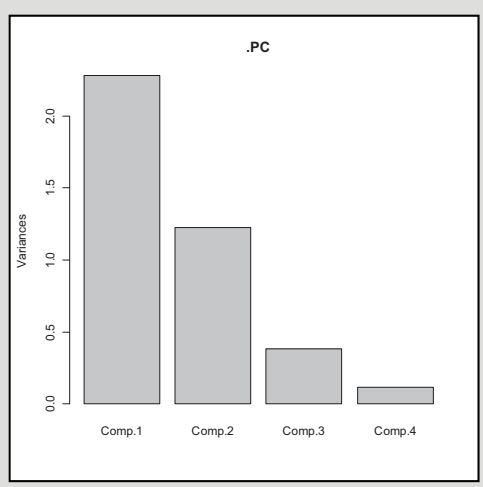
CUADRO XIII.1. *(Continuación)*[R]



Paso 10. En el cuadro «Número de componentes a retener», movemos el indicador a 2.



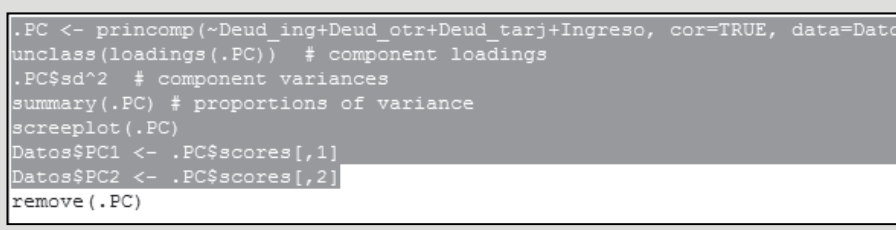
Paso 11. En primer lugar el gráfico de sedimentos muestra la forma en la que decrecen los autovalores o varianzas de las componentes. Las dos últimas componentes tienen una varianza muy pequeña.



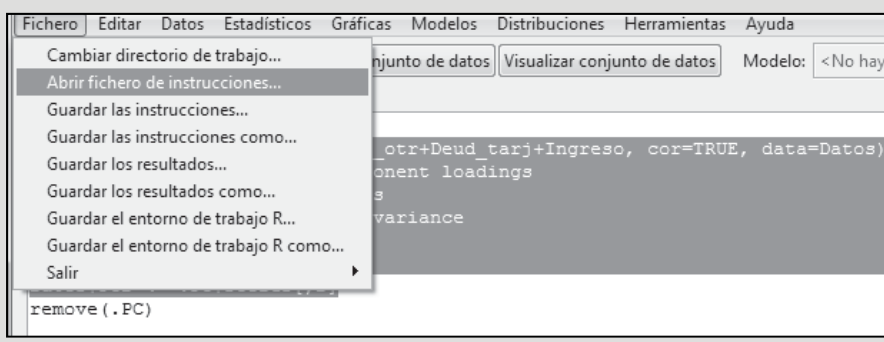
CUADRO XIII.1. *(Continuación)*[R]

Paso 12. Para facilitar la interpretación construiremos el gráfico de saturaciones. Para ello, se copian las instrucciones del Análisis de Componentes Principales, que se muestran a continuación y en la siguiente ventana.

```
.PC <- princomp(~Deud_ing+Deud_otr+Deud_tarj+Ingreso, cor=TRUE, data=Datos)
unclass(loadings(.PC))  # component loadings
.PC$sd^2  # component variances
summary(.PC) # proportions of variance
screeplot(.PC)
Datos$PC1 <- .PC$scores[,1]
Datos$PC2 <- .PC$scores[,2]
```



Paso 13. A continuación se carga el archivo de instrucciones **Cuadro XIII.1.R** seleccionando «Fichero» ⇨ «Abrir fichero de instrucciones…» y no se guarda la salida.



Paso 14. Se pegan las instrucciones que se copiaron en el paso 12, arriba de las instrucciones que se importaron del archivo **Cuadro XIII.1.R**. Las instrucciones del archivo **Cuadro XIII.1.R**, que se muestran a continuación, permiten realizar el gráfico de saturaciones. «Plot» (hace el gráfico) y «text» (añade al gráfico las etiquetas o nombres de variable).

```
plot(.PC$loadings, type="n")
text(.PC$loadings, labels=row.names(.PC$loadings), cex=.7)
```

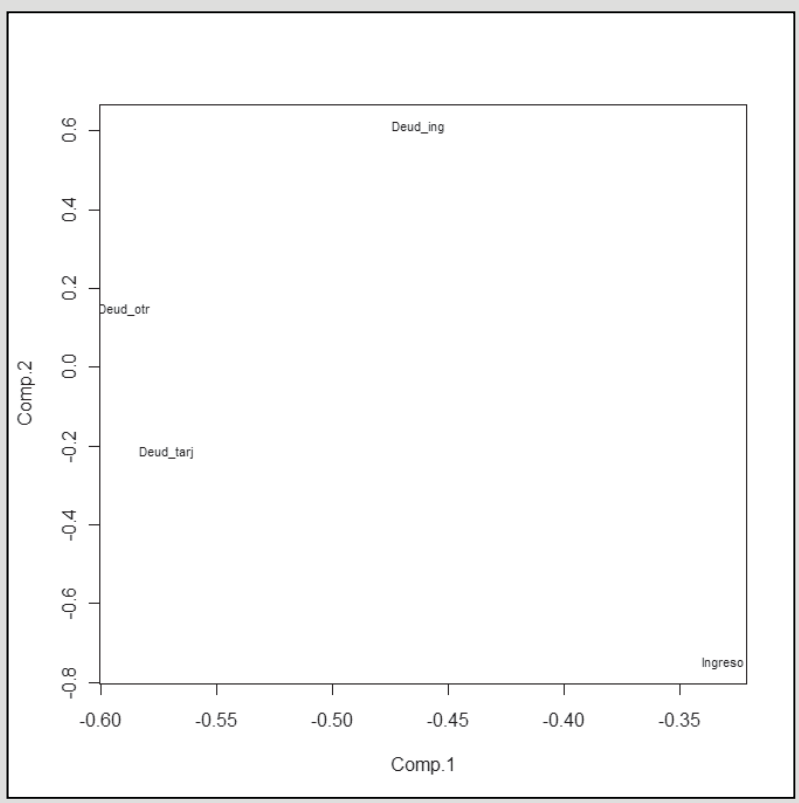
CUADRO XIII.1. *(Continuación)*[R]

Por último, se selecciona todo y se pulsa «Ejecutar»

```
.PC <- princomp(~Deud_ing+Deud_otr+Deud_tarj+Ingreso, cor=TRUE, data=Datos)
unclass(loadings(.PC))  # component loadings
.PC$sd^2  # component variances
summary(.PC) # proportions of variance
screeplot(.PC)
Datos$PC1 <- .PC$scores[,1]
Datos$PC2 <- .PC$scores[,2]
plot(.PC$loadings, type="n")
text(.PC$loadings, labels=row.names(.PC$loadings), cex=.7)
```

Ventana de resultados | Ejecutar

Paso 15. El gráfico resultante –que aparece en la ventana principal de R– se muestra a continuación.



Las variables más relacionadas con la primera componente son las que se alejan del cero en sentido horizontal: «Deud_otr» y «Deud_tarj». Con la segunda componente las que se alejan en sentido vertical son «Deud_ing» e «Ingreso». Esto permite asignar significados aproximados a las componentes, en función de las variables más relacionadas con cada una de ellas.

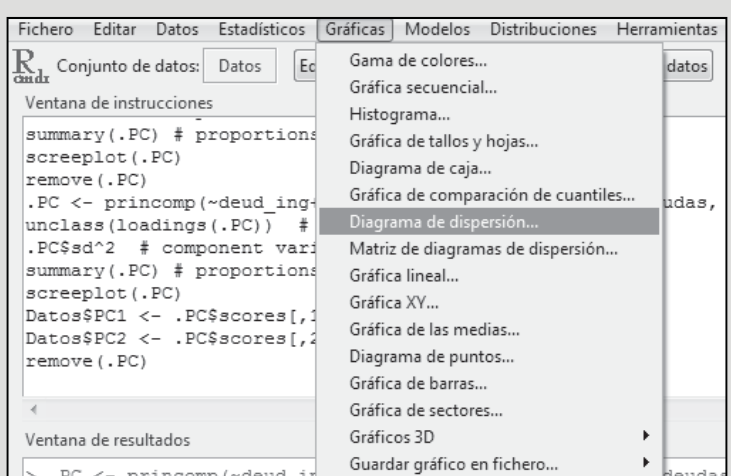
CUADRO XIII.1. *(Continuación)*[R]

Paso 16. En la ventana de resultados de R Commander encontramos las saturaciones o cargas, que expresan la combinación de variables originales que define cada una de las cuatro componentes principales generadas. Permiten averiguar cuales son las variables más relacionadas con cada componente.

```
> unclass(loadings(.PC))  # component loadings
             Comp.1     Comp.2     Comp.3     Comp.4
Deud_ing  -0.4630440  0.6114054 -0.1300850  0.6283723
Deud_otr  -0.5902370  0.1470053  0.6596690 -0.4414143
Deud_tarj -0.5720235 -0.2130375 -0.7054145 -0.3602699
Ingreso   -0.3316620 -0.7477887  0.2242879  0.5296294

> .PC$sd^2  # component variances
   Comp.1    Comp.2    Comp.3    Comp.4
2.2277452 1.2552330 0.3996087 0.1174131

> summary(.PC) # proportions of variance
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4
Standard deviation     1.4925633 1.1203718 0.63214608 0.34265600
Proportion of Variance 0.5569363 0.3138083 0.09990217 0.02935328
Cumulative Proportion  0.5569363 0.8707445 0.97064672 1.00000000
```

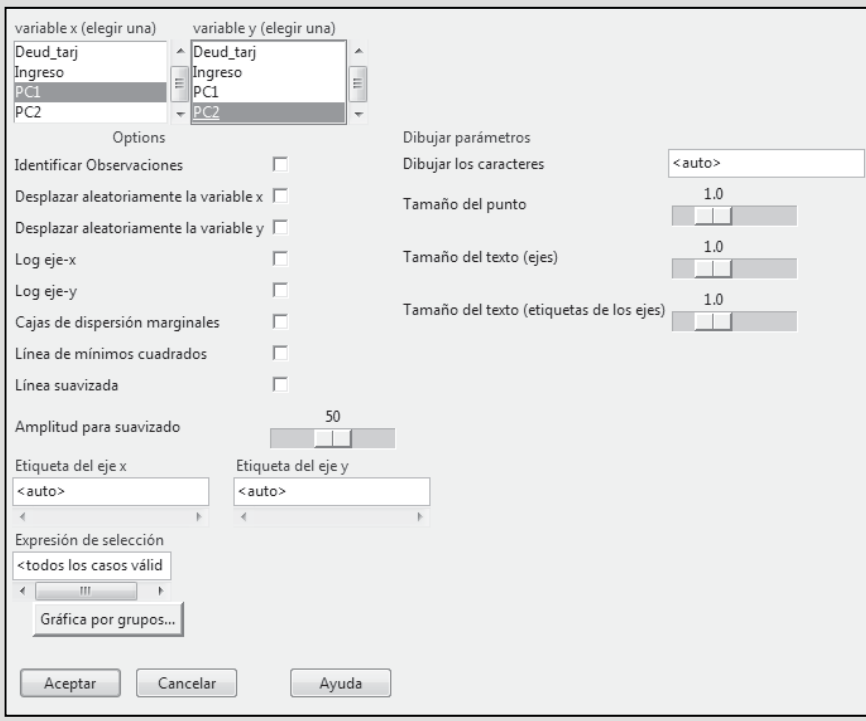
A continuación se muestran los autovalores o varianzas: 2,2277 para la primera componente y 1,2552 para la segunda. Más abajo aparecen expresadas en proporción del total: 55,69% la primera y 31,38% la segunda. Entre la primera y la segunda componente explican el 87,07% de la variabilidad total de los datos.

Paso 17. Se han añadido dos variables nuevas a nuestro conjunto de datos, las puntuaciones de las dos primeras componentes principales para cada caso o elemento fila. Representaremos gráficamente estos valores, para lo cual elegimos la opción «Gráficas» ⇨ «Diagrama de dispersión…» en el menú de R Commander.

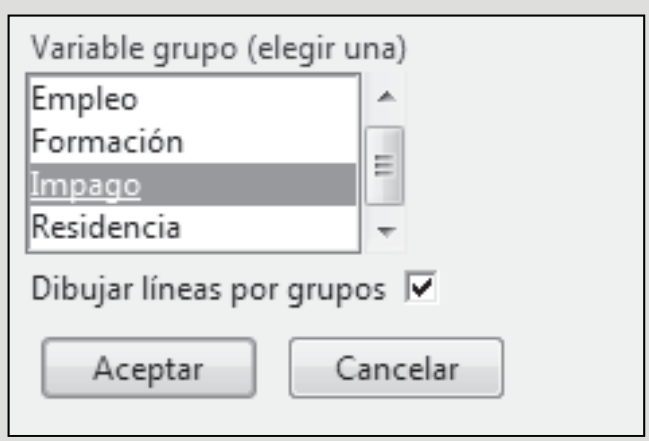


CUADRO XIII.1. *(Continuación)*[R]

Paso 18. En el cuadro de diálogo que aparece elegimos las variables a representar: «PC1» en el eje horizontal «variable x (elegir una)» y «PC2» en el eje vertical «variable y (elegir una)». Suprimimos las marcas por defecto.

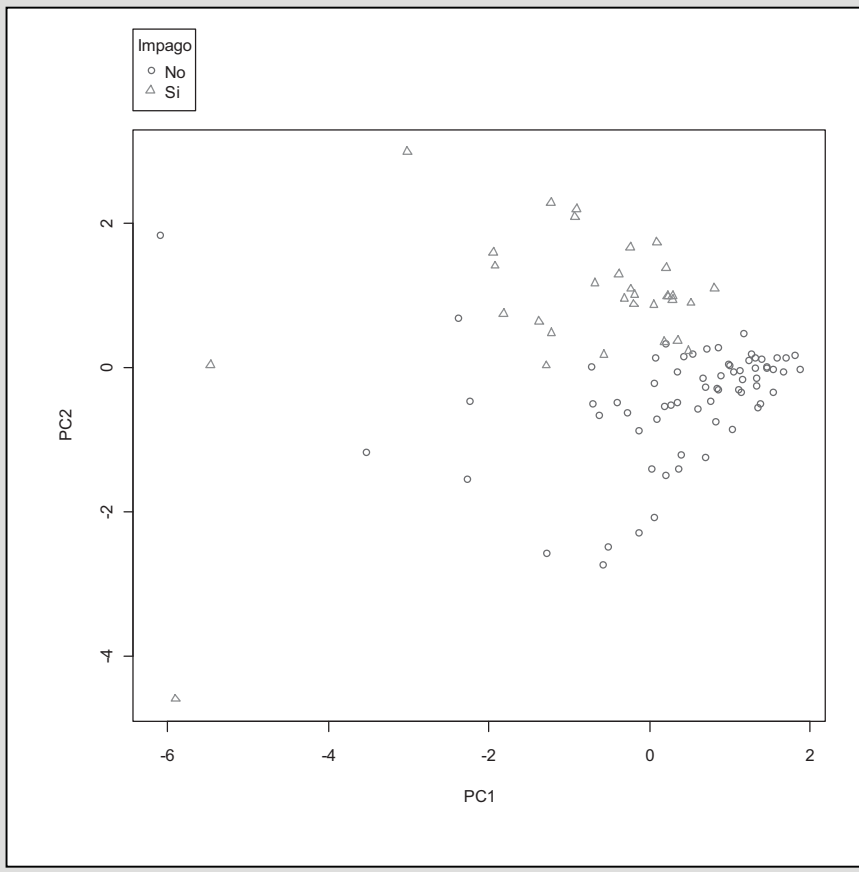


Paso 19. Al pulsar «Gráfica por grupos» aparece la siguiente ventana donde seleccionamos la variable «impago». Esto permitirá identificar con una marca los clientes con y sin impagos.



CUADRO XIII.1. *(Continuación)*[R]

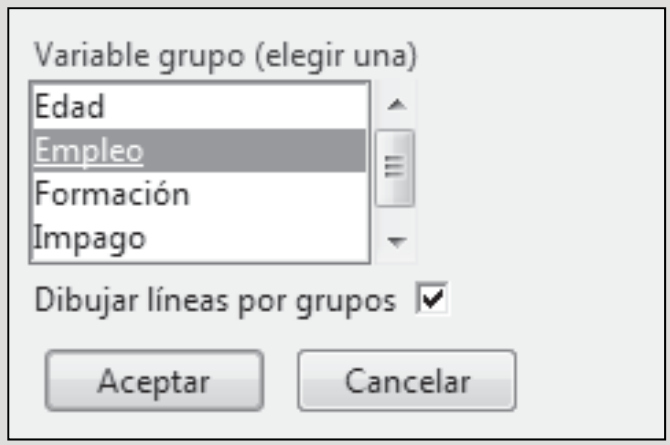
Paso 20. En el gráfico resultante podemos apreciar como los clientes con impagos forman un grupo en la parte superior. Corresponden por lo tanto a los valores altos de la segunda componente principal.



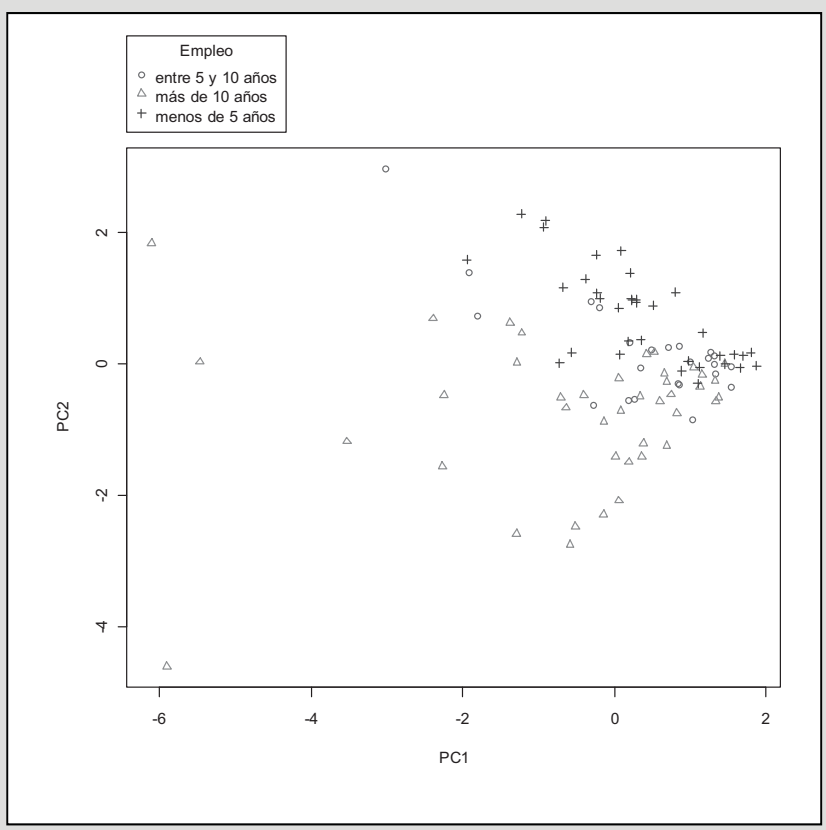
La segunda componente está asociada fundamentalmente a las variables «ingreso» y «deuda_ingreso», cuyos coeficientes o saturaciones son respectivamente (ver pasos 15 y 16) -0,747 y 0,611. Esto nos dice que los clientes con impagos tienen ingresos bajos (ya que el coeficiente es negativo) y una alta relación deudas/ingresos. El gráfico nos permite por lo tanto describir nuestro conjunto de datos en términos de la situación financiera, relacionando la información disponible.

Paso 21. Repetimos los pasos 18 a 20 pero esta vez seleccionamos la variable «Empleo» como se muestra en la siguiente ventana.

CUADRO XIII.1. *(Continuación)*[R]



Paso 22. La gráfica muestra que los clientes con impagos –los valores altos en el eje 2– son lo que llevan menos tiempo trabajando, mientras que el número de clientes con impagos es menor en las personas que llevan más tiempo con empleo.

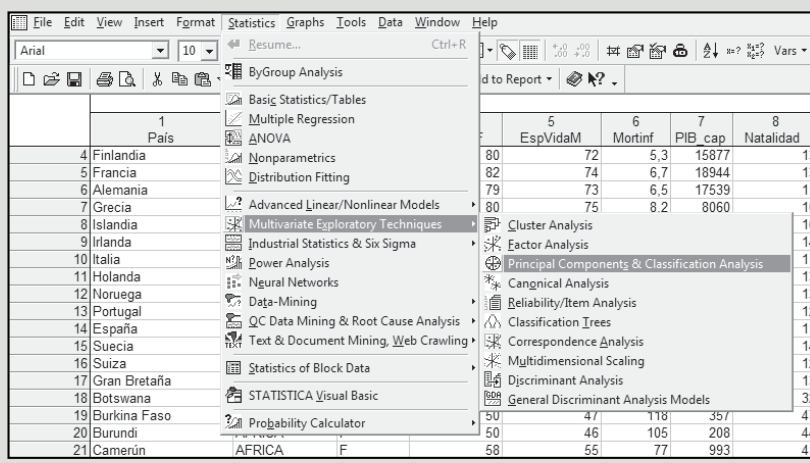


CUADRO XIII.1. *(Continuación)*[STATISTICA]

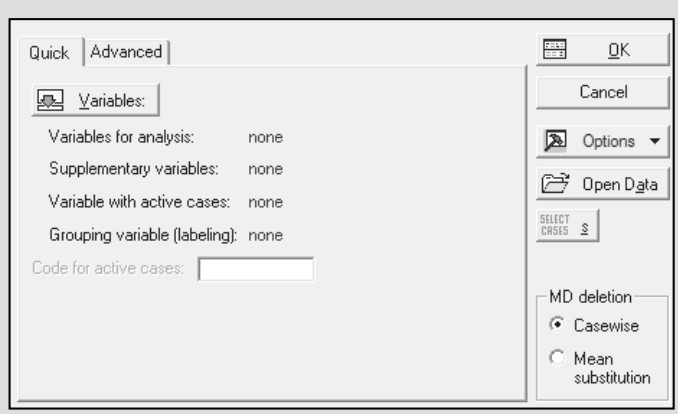
APLICACIÓN CON STATISTICA.

EJEMPLO. Utilizaremos los datos contenidos en el archivo **Cuadro XIII.1.sta**. Se trata de un conjunto de 57 países de Europa, África y América. Las variables indican la esperanza de vida masculina y femenina al nacer (en años de vida), las tasas de mortalidad, mortalidad infantil, natalidad, y fertilidad, el Producto Interior Bruto per cápita (en miles de dólares anuales), y la tasa de alfabetización de hombres y mujeres (en porcentaje). Todas estas variables están relacionadas entre sí (correlacionadas), ya que se asocian al grado de desarrollo y, por ello, podemos intentar resumir la información de las nueve variables en dos componentes, con el fin de representar los países en un gráfico.

Paso 1. Una vez abierto el archivo de datos con STATISTICA, para aplicar el Análisis de Componentes Principales elegimos en el menú principal las opciones «Statistics» ⇨ «Multivariate Exploratory Techniques» ⇨ «Principal Components & Classification Analysis».

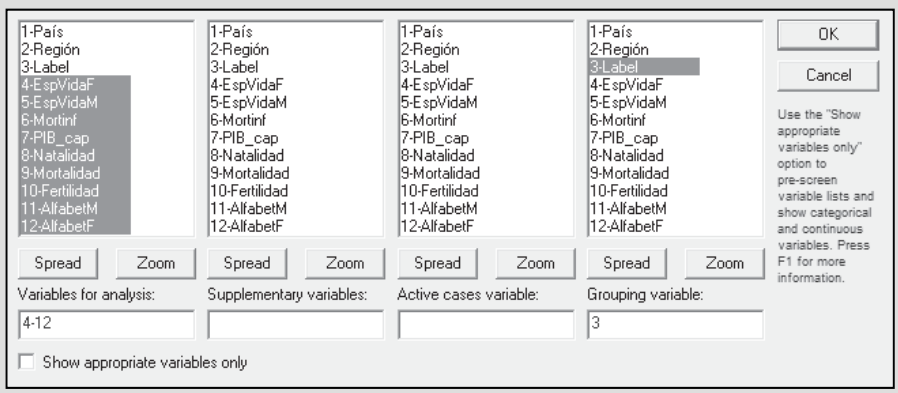


Paso 2. En la siguiente ventana pulsamos en «Variables».

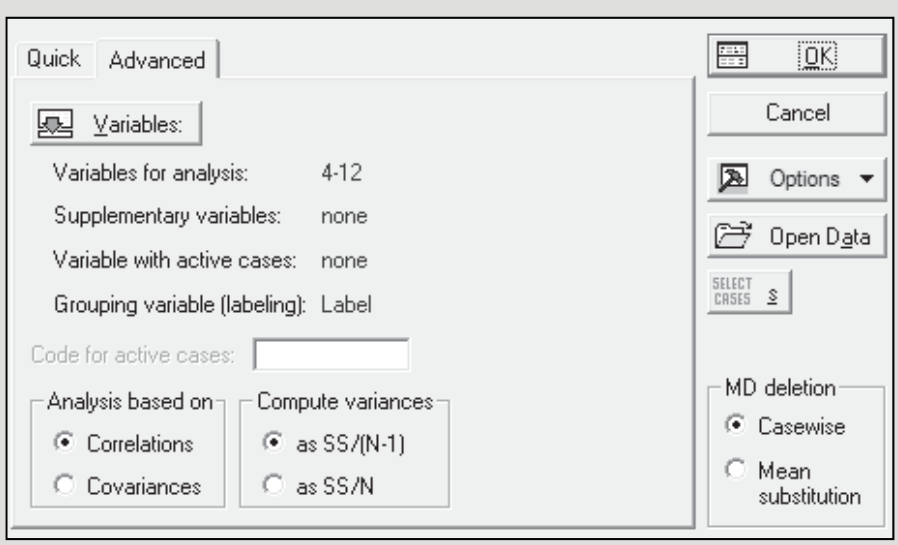


CUADRO XIII.1. *(Continuación)* *STATISTICA*

Paso 3. Se seleccionan las variables –todas las variables cuantitativas de nuestros datos– en la primera columna «Variables for analysis:». Elegimos también la variable «Label» en la última columna «Grouping variable:», que utilizaremos en el gráfico final.

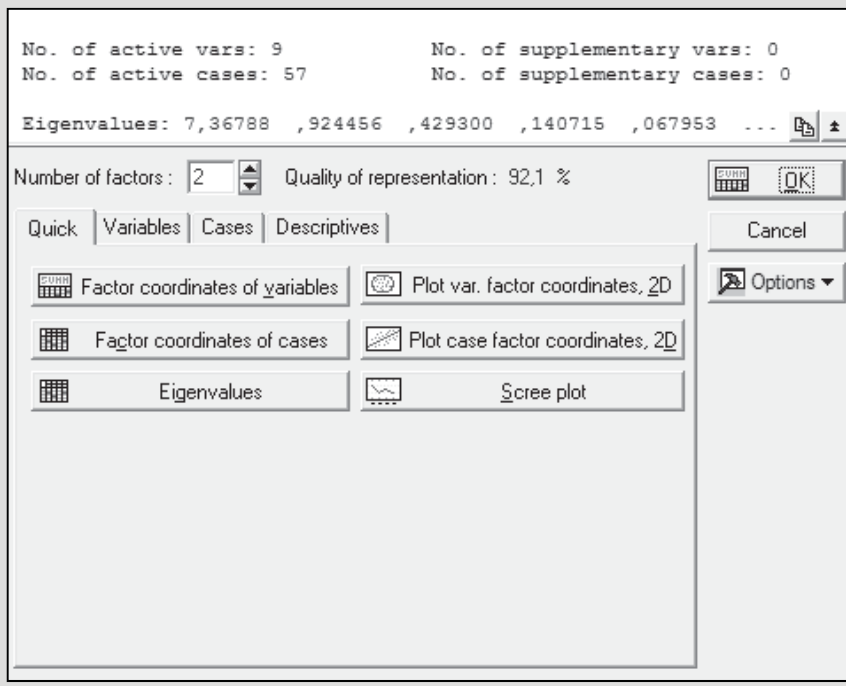


Paso 4. En la pestaña «Advanced» marcamos «Analysis based on: Correlations», dado que nuestras variables son heterogéneas (si la unidad de medida fuese igual para todas emplearíamos la matriz de covarianzas), y «Compute variances: as SS/N», ya que nuestros datos no son una muestra, sino todo el conjunto de países que queremos analizar. Si se trata de una muestra representativa es preferible utilizar las cuasivarianzas «SS/(N-1)» como estimación de las varianzas poblacionales.



CUADRO XIII.1. *(Continuación)*[STATISTICA]

Paso 5. Aparece una ventana con algunos resultados de interés: Número de casos y variables, y los autovalores (Eigenvalues): 7,367, 0,924 0,429, 0,1407, etc. Estos autovalores son las varianzas de las componentes principales. Debemos decidir cuantas conservamos, moviendo el indicador «Number of factors» a 2. Observaremos como a la derecha de este número aparece el porcentaje de varianza explicada «Quality or representation», el 92,1%.

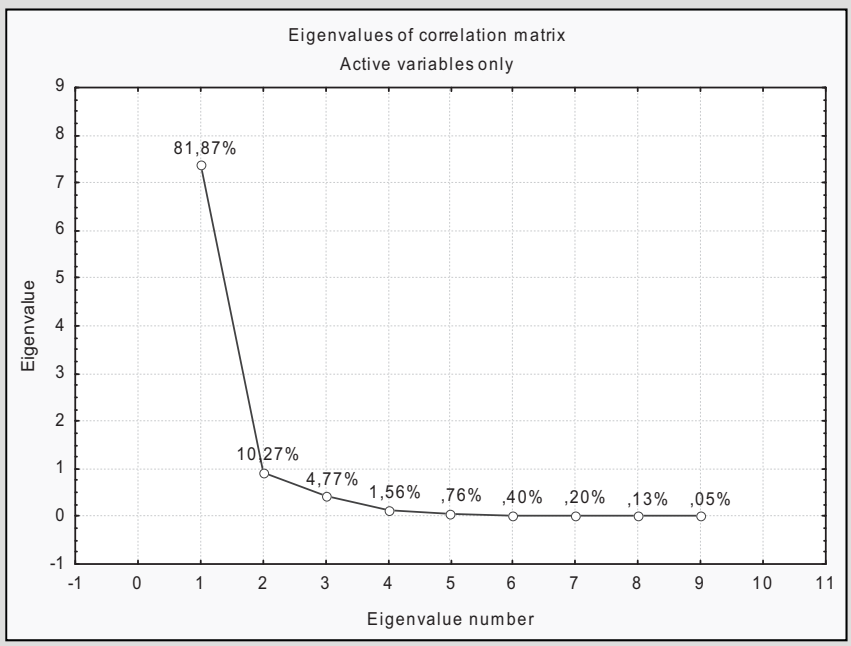


Paso 6. En la pestaña «Quick» se pulsa en «Eigenvalues» y aparecen todos los autovalores, expresados también como porcentaje de la varianza total y como porcentaje acumulado.

	Eigenvalues of correlation matrix, and related statistics (Cuadro XIII.1.sta) Active variables only			
Value number	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	7,367877	81,86530	7,367877	81,8653
2	0,924456	10,27173	8,292333	92,1370
3	0,429300	4,77000	8,721633	96,9070
4	0,140715	1,56350	8,862348	98,4705
5	0,067953	0,75503	8,930301	99,2256
6	0,035612	0,39569	8,965913	99,6213
7	0,017772	0,19746	8,983684	99,8187
8	0,011893	0,13214	8,995577	99,9509
9	0,004423	0,04914	9,000000	100,0000

CUADRO XIII.1. *(Continuación)*^{STATISTICA}

Paso 7. Al pulsar «Scree plot» obtenemos el gráfico de sedimentos. Vemos como el primer autovalor es mucho más importante que el segundo y la curva muestra un codo a partir de la segunda componente. Esto sugiere que una sola componente o dos serán suficientes para conservar una parte importante de la información.

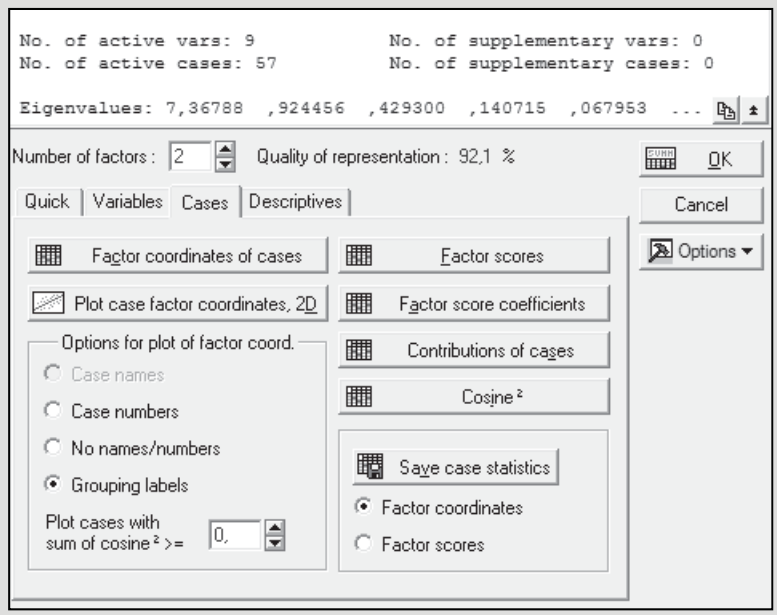


Paso 8. En «Factor coordinates of variables», o saturaciones, muestra los coeficientes de las combinaciones lineales que definen a cada una de las componentes principales. La primera componente tiene saturaciones altas en todas las variables, lo que sugiere una idea global de «Desarrollo». La segunda por el contrario, solo tiene dos saturaciones importantes: «PIB_cap» y «Mortalidad». Estas dos variables podrían explicar algunas diferencias importantes entre los países más allá de su nivel de desarrollo.

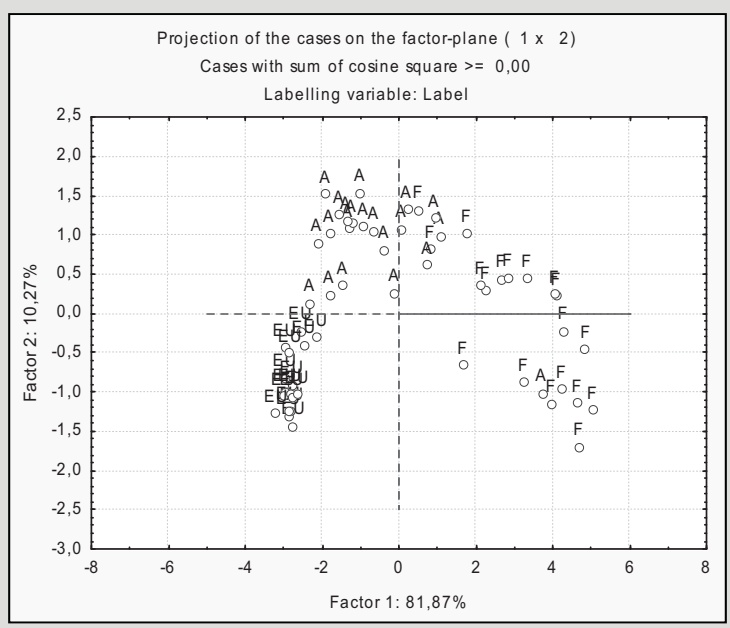
	Factor coordinates of the variables, based on correlations (Cuadro XIII.1.sta)	
Variable	Factor 1	Factor 2
EspVidaF	-0,976553	0,126720
EspVidaM	-0,967163	0,146298
Mortinf	0,973451	-0,054415
PIB_cap	-0,728551	-0,594224
Natalidad	0,959254	0,180782
Mortalidad	0,692785	-0,693307
Fertilidad	0,963790	0,072907
AlfabetM	-0,896386	-0,107189
AlfabetF	-0,931615	-0,027735

CUADRO XIII.1. *(Continuación)*[STATISTICA]

Paso 9. En la pestaña «Cases» marcamos dentro de las opciones de gráfico la opción «Grouping labels».



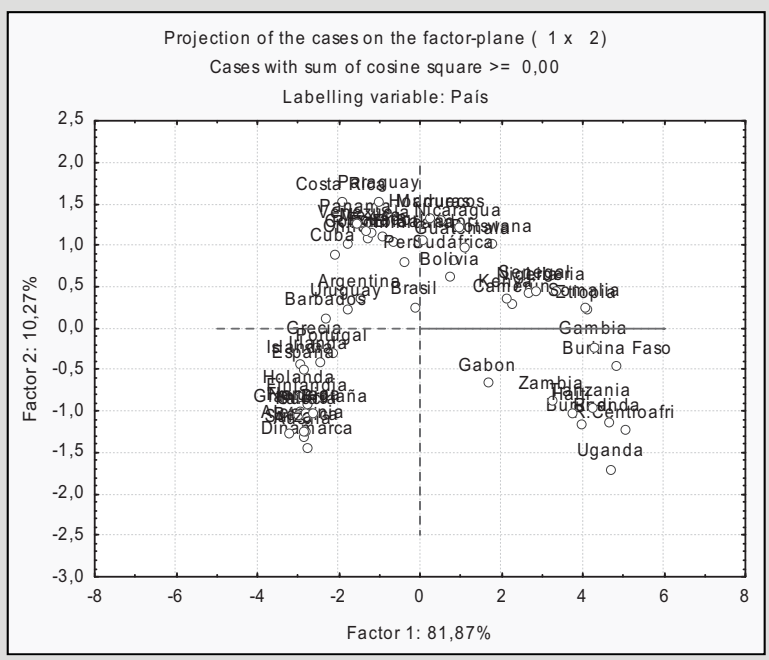
Paso 10. En la ventana del paso 5 pulsamos en «Plot case factor coordinates. 2D» y obtenemos el gráfico de los países en el plano de las dos primeras componentes.



CUADRO XIII.1. *(Continuación)*[STATISTICA]

La etiqueta de región (variable «Label») permite distinguir los países de cada continente. Observamos como los países del mismo continente están próximos entre sí, formando tres grupos que se distinguen claramente. Esto sugiere que los países próximos tienen un desarrollo económico similar.

Paso 11. En el paso 3, si en vez de seleccionar «Label» se selecciona «País», se obtendría el siguiente gráfico, donde se pueden ver las agrupaciones de los diferentes países. Ahora observamos que los países más desarrollados (Dinamarca, España, Finlandia, etc.) están a la izquierda, y los menos avanzados (Gambia, Uganda, Somalia, etc.) a la derecha del gráfico. La primera componente parece explicar, por lo tanto, el nivel de desarrollo.



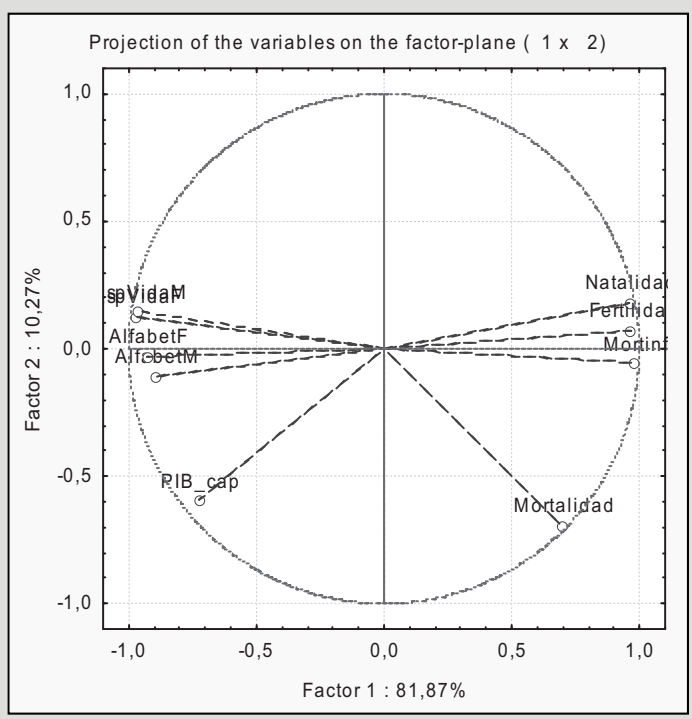
Paso 12. Para poder ver exactamente la posición de los países en los ejes I y II de componentes principales, en la ventana del paso 5 pulsamos en «Factor coordinates of cases».

	Factor coordinates of cases, based on correlations (Cuadro XIII.1.sta) Labelling variable: País		
Case	Factor 1	Factor 2	País
1	-2,86835	-1,29461	Austria
2	-2,80696	-1,24151	Bélgica
3	-2,76307	-1,42899	Dinamarca

CUADRO XIII.1. *(Continuación)*[STATISTICA]

Paso 13. Para determinar cuáles son las variables que explican los ejes de componentes principales, en la ventana del paso 5 pulsamos en «Plot var. factor coordinates. 2D». En el gráfico se observa claramente que en el extremo izquierdo del eje I están las variables alfabetización de hombres y mujeres y esperanza de vida también en hombres y mujeres. Valores altos de estas variables se asocian con países desarrollados. Por el contrario, en el extremo derecho del eje I están variables como alta natalidad, alta fertilidad y alta mortalidad infantil, las cuales se asocian con países no desarrollados.

¿Cuál es el significado de la segunda componente? El gráfico nos recuerda que sólo explica el 10,27% de la variabilidad, por lo que es mucho menos importante que la primera. Si recordamos que las variables con mayor saturación eran «PIB_cap» y «Mortalidad» podemos interpretar las diferencias en sentido vertical en función de esta última variable: los países más atrasados y los más adelantados comparten altas tasas de mortalidad (Dinamarca, Suecia o Gran Bretaña tienen tasas de mortalidad similares a Kenia, Camerún o Nigeria, que duplican las de otros países con nivel de desarrollo intermedio). En los más atrasados las altas tasas se deben al escaso nivel socio-sanitario, en los más adelantados se deben a causas de mortalidad ligadas al desarrollo: Accidentes de tráfico, contaminación, o enfermedades coronarias, entre otras. El gráfico del paso 11 nos dice que el camino del desarrollo (desde Uganda hasta Dinamarca) no es recto, sino curvo y bidimensional.

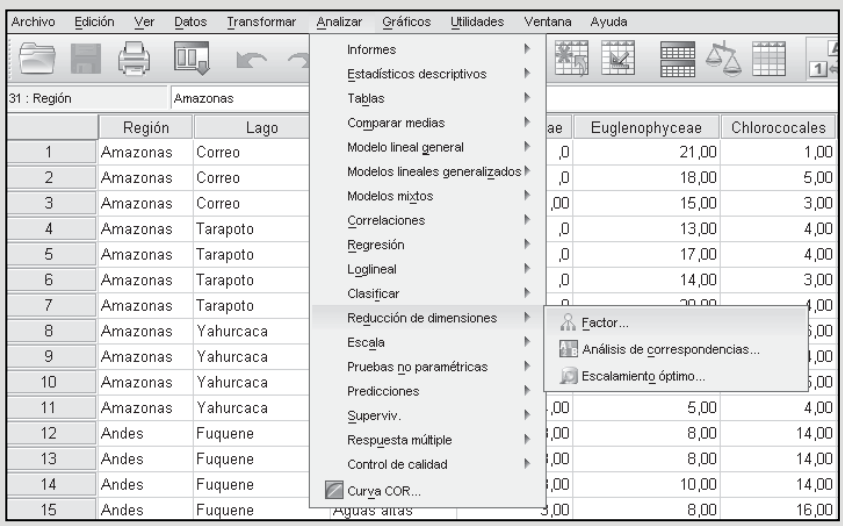


CUADRO XIII.1. *(Continuación)*[SPSS]

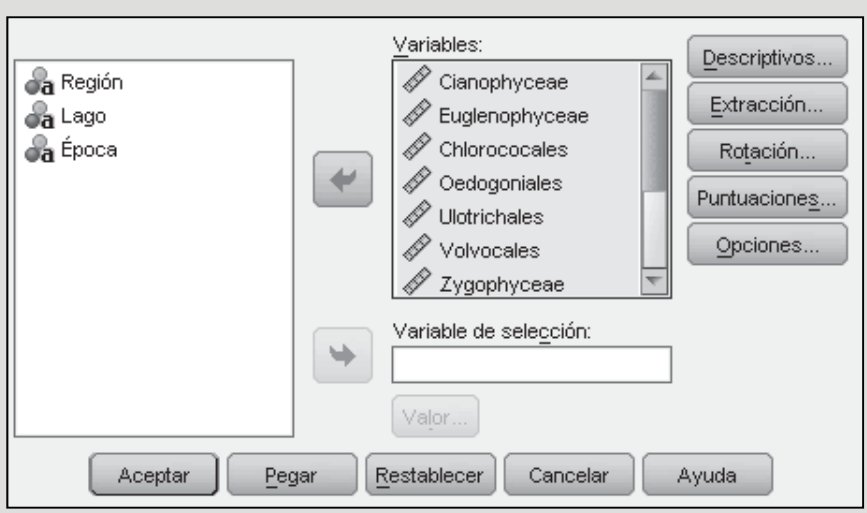
Aplicación con SPSS.

EJEMPLO. En el archivo **Cuadro XIII.1.sav** hay datos del número de especies, de diferentes grupos de algas del fitoplancton, presentes en lagos de diferentes regiones geográficas y en dos épocas distintas del año. El objetivo es determinar si es posible explicar la composición de especies del fitoplancton en base a la región y la época del año.

Paso 1. Se selecciona «Analizar» ⇨ «Reducción de dimensiones» ⇨ «Factor…».

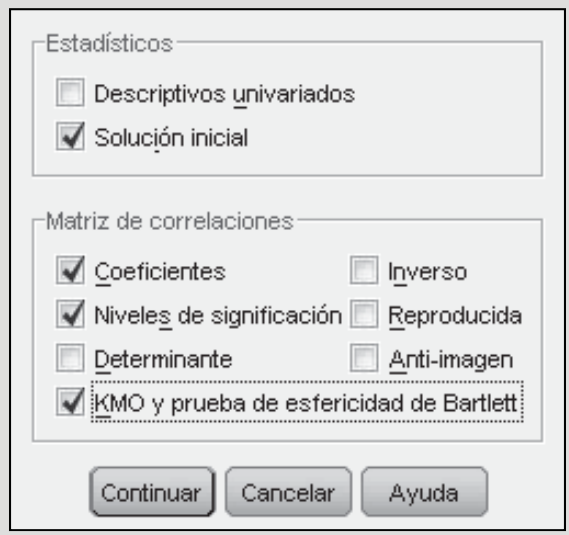


Paso 2. Se seleccionan todas las variables de interés, en este caso todos los grupos de fitoplancton. Se pulsa en «Descriptivos…».

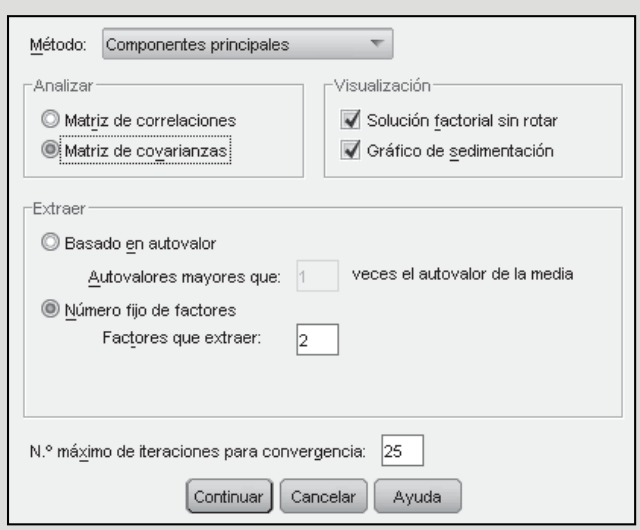


CUADRO XIII.1. *(Continuación)*[SPSS]

Paso 3. En la ventana que aparece se selecciona «Solución inicial», «Coeficientes», «Nivel de significación» y «KMO y prueba de esfericidad de Bartlett».

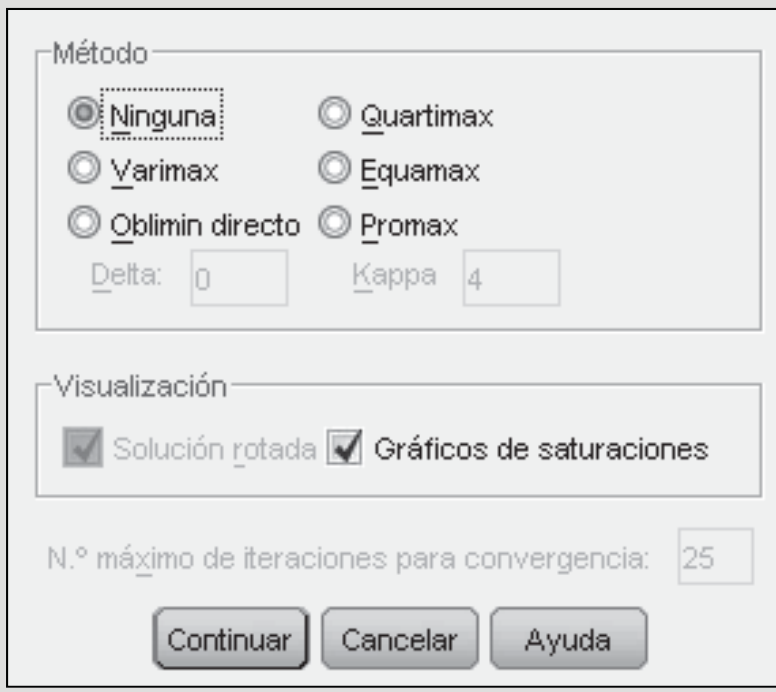


Paso 4. En la ventana del paso 2 se pulsa en «Extracción…». En el apartado «Método» mantenemos la opción por defecto, «Componentes principales», en «Analizar» seleccionamos «Matriz de covarianzas» -ya que nuestras variables son homogéneas, es decir, están en las mismas unidades-, en «Visualización» dejamos la opción «Solución factorial sin rotar» y marcamos «Gráfico de sedimentación». Por último, en «Extraer» marcamos «Número fijo de factores» y escribimos el número que nos interesa: 2 factores.

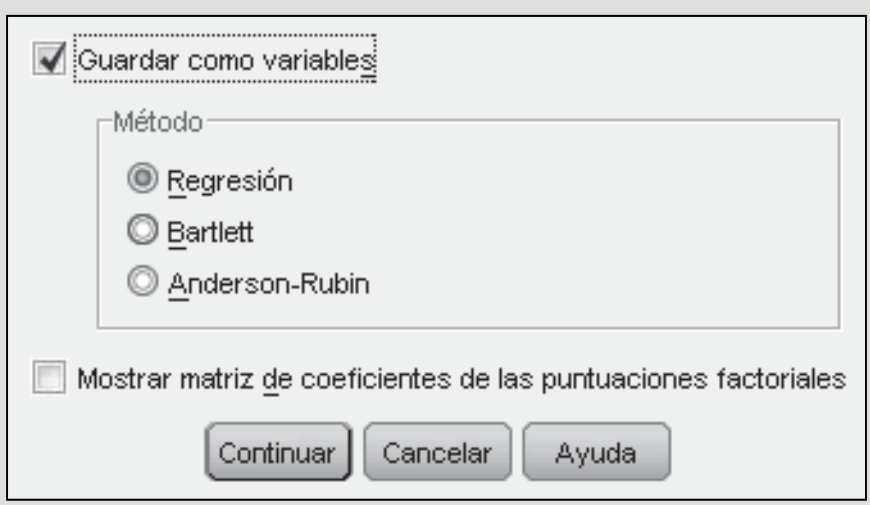


CUADRO XIII.1. *(Continuación)*[SPSS]

Paso 5. En la ventana del paso 2 se pulsa en «Rotación…». «En Método» se deja por defecto «Ninguna» y marcamos «Gráfico de saturaciones».

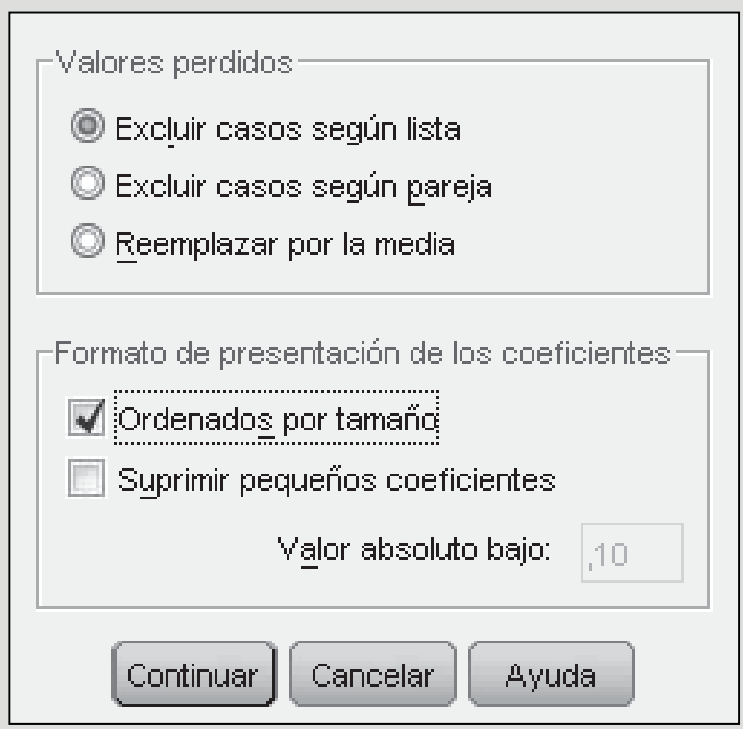


Paso 6. En el cuadro de diálogo de «Puntuaciones…» marcamos «Guardar como variables», con la opción por defecto «Regresión».



CUADRO XIII.1. *(Continuación)*[SPSS]

Paso 7. En el cuadro de diálogo de «Opciones…» mantenemos la opción por defecto «Excluir casos según lista», ya que no tenemos valores perdidos. Si nuestros datos contienen valores perdidos, sería mejor utilizar «Excluir casos según pareja», procedimiento menos eficiente y más lento de cálculo, pero que utiliza toda la información disponible. Marcamos «Ordenados por tamaño», para facilitar la interpretación de las componentes principales. En el caso de tener muchas variables y más de dos componentes puede ser recomendable también utilizar la opción «Suprimir pequeños coeficientes» para eliminar de los resultados los valores pequeños (por ejemplo < 0,4), que estorban en la lectura de la tabla.



Paso 8. El test de esfericidad de Bartlett (p < 0,001) indica que existe correlación.

KMO y prueba de Bartlett[a]

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,563
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	271,138
	gl	66
	Sig.	,000

a. Basado en correlaciones

CUADRO XIII.1. *(Continuación)*[SPSS]

Paso 9. Sin embargo, las correlaciones -de las cuales solo se muestra una parte en la siguiente tabla- muestran que hay muchos valores con probabilidad mayor que 0,05 lo que indica que muchos coeficientes de correlación parecen ser nulos. Es necesario recordar que cuanto más alta sea la correlación entre las variables, mayor es la proporción de varianza explicada por las primeras componentes y, por lo tanto, más pequeña será la pérdida de información al conservarlas prescindiendo de las restantes. Por tanto, la existencia de muchas correlaciones no significativas no favorece a que las primeras Componentes Principales expliquen una alta proporción de la varianza observada.

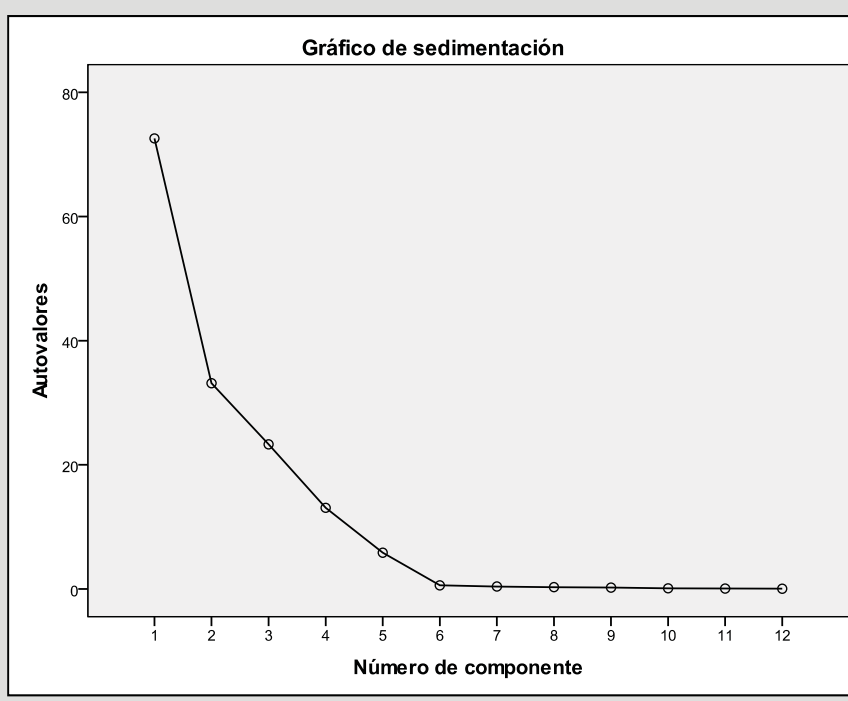
		Cianophyceae	Euglenophyceae	Chlorococales
Correlación	Cianophyceae	1,000	-,359	,567
	Euglenophyceae	-,359	1,000	-,185
	Chlorococales	,567	-,185	1,000
	Oedogoniales	-,249	,155	-,221
	Ulotrichales	,165	,084	,112
	Volvocales	,235	,365	,197
	Zygophyceae	,033	-,139	,373
	Bacillariophyceae	-,089	,433	,055
	Crysophyceae	-,063	,240	,228
	Cryptophyceae	-,404	,676	-,266
	Dinophyceae	-,145	-,352	,106
	Xanthophyceae	,518	,160	,089
Sig. (Unilateral)	Cianophyceae		,002	,000
	Euglenophyceae	,002		,078
	Chlorococales	,000	,078	
	Oedogoniales	,027	,119	,045
	Ulotrichales	,104	,262	,198
	Volvocales	,035	,002	,066
	Zygophyceae	,401	,145	,002
	Bacillariophyceae	,249	,000	,338
	Crysophyceae	,316	,032	,040
	Cryptophyceae	,001	,000	,020
	Dinophyceae	,134	,003	,210
	Xanthophyceae	,000	,111	,249

Paso 10. La siguiente tabla muestra los autovalores, los cuales indican la varianza de cada componente, que aparece también expresada en porcentaje de la variabilidad total. La primera componente resume el 48,48% de la variabilidad y la segunda el 22,14%. Entre ambas solamente el 70,629% del total.

CUADRO XIII.1. *(Continuación)*[SPSS]

	Componente	Autovalores iniciales[a]			Sumas de las saturaciones al cuadrado de la extracción		
		Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
Bruta	1	72,575	48,484	48,484	72,575	48,484	48,484
	2	33,149	22,145	70,629	33,149	22,145	70,629
	3	23,300	15,566	86,194			
	4	13,071	8,732	94,926			
	5	5,845	3,905	98,831			
	6	,587	,392	99,223			
	7	,401	,268	99,491			
	8	,296	,198	99,689			
	9	,226	,151	99,840			
	10	,114	,076	99,916			
	11	,076	,051	99,967			
	12	,050	,033	100,000			

Paso 11. El gráfico de sedimentos muestra también esta diferente importancia de cada componente en función de la variabilidad explicada. El hecho de que la curva se aleje del origen de coordenadas y no muestre una pendiente inicial muy acusada indica unos datos con escasa correlación y, por consiguiente, como se mencionó anteriormente poca eficacia en la aplicación del Análisis de Componentes Principales.



CUADRO XIII.1. *(Continuación)*[SPSS]

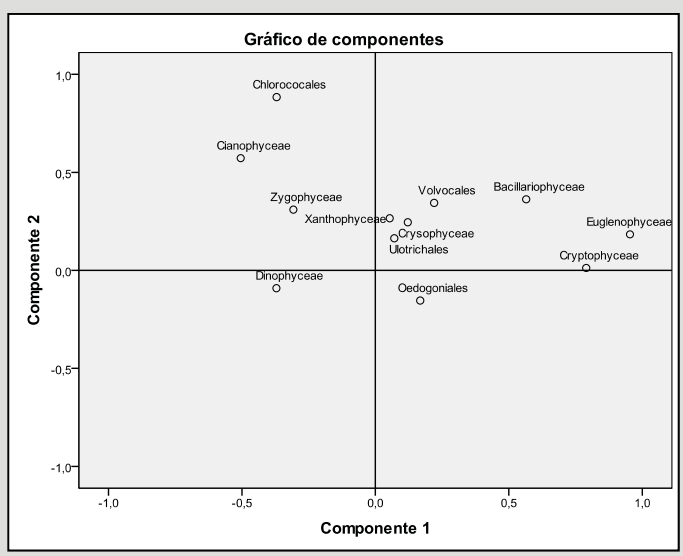
Paso 12. El siguiente resultado es la matriz de coeficientes o saturaciones, que están ordenados por componentes. Hay que fijarse en la matriz reescalada. La primera componente aparece claramente asociada a Euglenophyceae, Cryptophyceae, Bacillariophyceae y Cianophyceae. La segunda componente tiene principalmente asociadas Chlorococales y Cianophyceae.

Matriz de componentes[a]

	Bruta		Reescalada	
	Componente		Componente	
	1	2	1	2
Euglenophyceae	7,350	1,415	,954	,184
Cryptophyceae	,842	,013	,790	,012
Bacillariophyceae	2,781	1,784	,565	,362
Dinophyceae	-,362	-,089	-,371	-,091
Oedogoniales	,059	-,055	,168	-,154
Chlorococales	-1,875	4,478	-,370	,884
Cianophyceae	-2,159	2,449	-,505	,572
Volvocales	,169	,265	,220	,344
Zygophyceae	-1,327	1,338	-,307	,310
Xanthophyceae	,036	,177	,054	,266
Crysophyceae	,047	,096	,121	,245
Ulotrichales	,018	,041	,071	,164

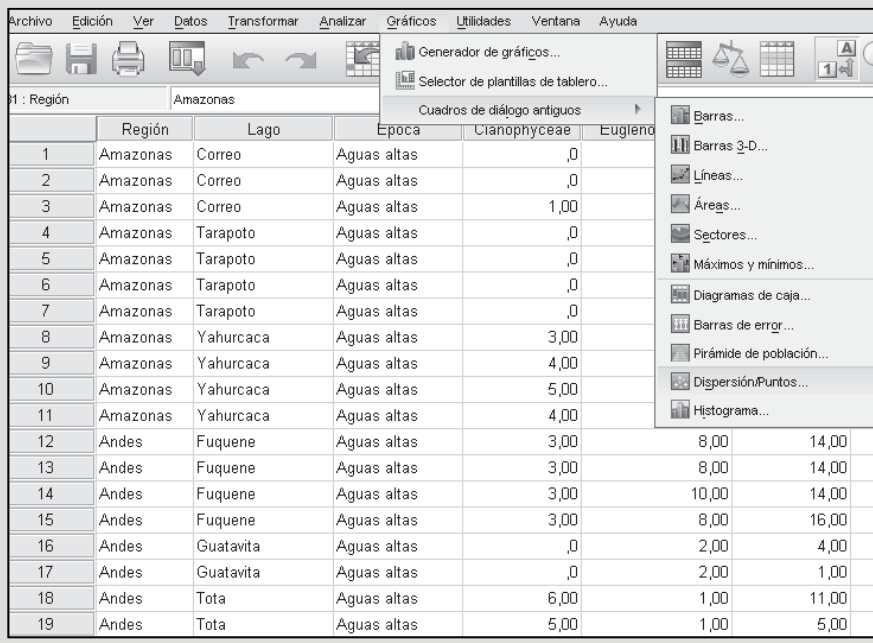
Método de extracción: Análisis de componentes principales.
a. 2 componentes extraídos

Paso 13. El gráfico de saturaciones representa los coeficientes de las funciones que definen cada componente principal, por lo que también se puede utilizar para su interpretación. Las variables más alejadas del origen en sentido horizontal son las que dan significado a la primera componente, y las que se alejan en sentido vertical a la segunda.

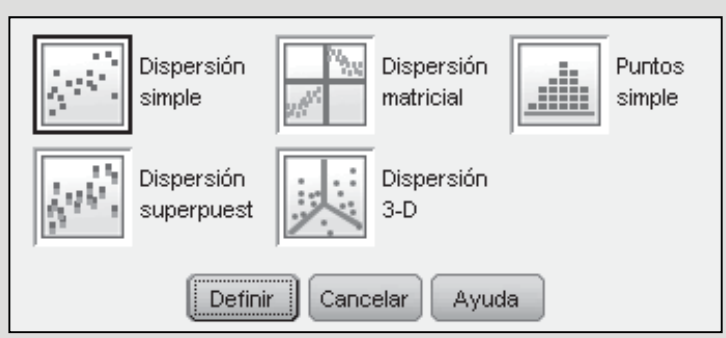


CUADRO XIII.1. *(Continuación)*[SPSS]

Paso 14. Probablemente el resultado de mayor interés en la aplicación de este método reside en la posibilidad de representar gráficamente los elementos de nuestro conjunto en el plano de las primeras componentes, que resumen la información de todas las variables utilizadas. Para eso le hemos pedido en el paso 6, que guardara las puntuaciones como nuevas variables. Se selecciona «Gráficos» ⇨ «Dispersión/Puntos…» ⇨ «Dispersión simple».

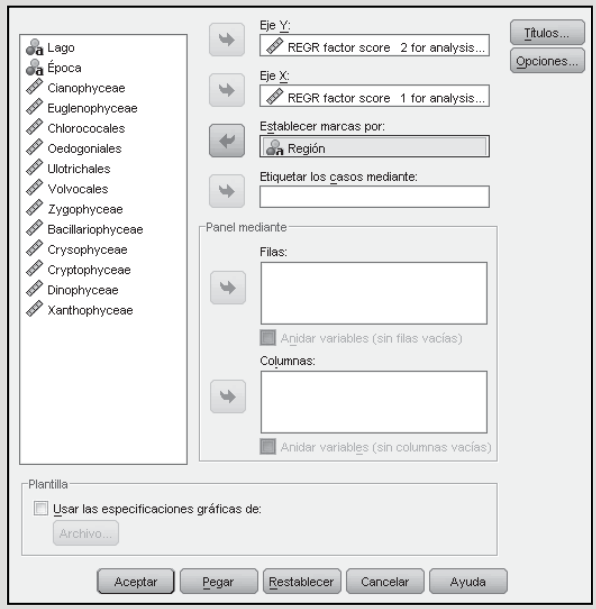


	Región	Lago	Época	Cianophyceae	Euglenophyceae	
1	Amazonas	Correo	Aguas altas	,0		
2	Amazonas	Correo	Aguas altas	,0		
3	Amazonas	Correo	Aguas altas	1,00		
4	Amazonas	Tarapoto	Aguas altas	,0		
5	Amazonas	Tarapoto	Aguas altas	,0		
6	Amazonas	Tarapoto	Aguas altas	,0		
7	Amazonas	Tarapoto	Aguas altas	,0		
8	Amazonas	Yahurcaca	Aguas altas	3,00		
9	Amazonas	Yahurcaca	Aguas altas	4,00		
10	Amazonas	Yahurcaca	Aguas altas	5,00		
11	Amazonas	Yahurcaca	Aguas altas	4,00		
12	Andes	Fuquene	Aguas altas	3,00	8,00	14,00
13	Andes	Fuquene	Aguas altas	3,00	8,00	14,00
14	Andes	Fuquene	Aguas altas	3,00	10,00	14,00
15	Andes	Fuquene	Aguas altas	3,00	8,00	16,00
16	Andes	Guatavita	Aguas altas	,0	2,00	4,00
17	Andes	Guatavita	Aguas altas	,0	2,00	1,00
18	Andes	Tota	Aguas altas	6,00	1,00	11,00
19	Andes	Tota	Aguas altas	5,00	1,00	5,00

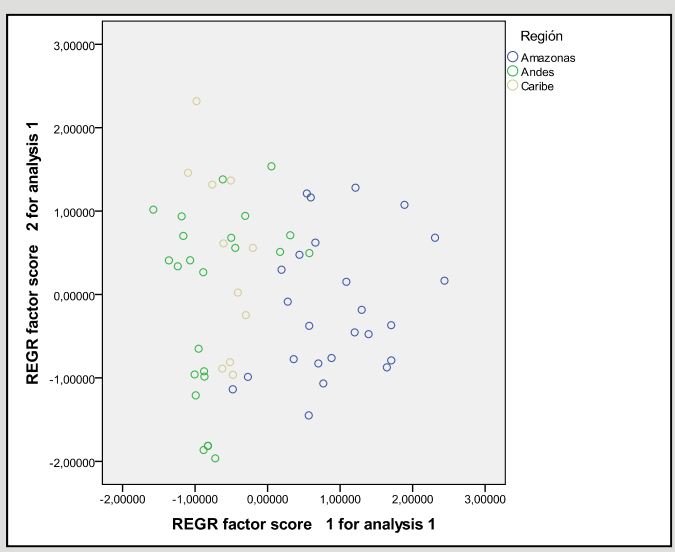


Paso 15. En el cuadro de diálogo que aparece elegimos las variables a representar: «REGR factor score 1» en el eje horizontal (eje *X*) y «REGR factor score 2» en el eje vertical (eje *Y*). En «Establecer marcas por:» introducimos la variable «Región», para diferenciar en función del área geográfica.

CUADRO XIII.1. *(Continuación)*[SPSS]



Paso 16. Se observa que la primera componente diferencia los lagos del Amazonas (asociados a altos valores de Euglenophyceae, Cryptophyceae, y Bacillariophyceae), de los lagos del Caribe y los Andes (con altos valores de Cianophyceae). La diferenciación de la segunda componente no es clara. En vez de utilizar «Región» se puede usar «Época» y observaremos que no existe una clara separación entre las dos épocas en el análisis. Por tanto, la componente temporal no es tan importante como la componente espacial, a la hora de explicar la variabilidad observada en nuestros datos.



XIII.2. Análisis Factorial

El Análisis Factorial utiliza un modelo inverso al del Análisis de Componentes Principales. Las variables observables y medibles, aquellas de las que tenemos datos, se expresan como combinación lineal de un conjunto reducido de factores comunes, establecidos a priori partiendo del conocimiento que tenemos de las variables. Así, podemos suponer que las diferentes variables que medimos en un conjunto de seres vivos (longitud, anchura, diámetro, etc.) son el resultado de un factor "genético" y un factor "ambiental"; o las diferentes variables obtenidas del análisis de los balances de las empresas (rentabilidad, liquidez, solvencia, crecimiento, etc.) son debidas a un factor de "coyuntura económica", otro de "gerencia" y un tercero de "estructura empresarial"; o bien que las diferentes manifestaciones de inteligencia en las personas (habilidad numérica, memoria, capacidad verbal, habilidad geométrica, capacidad de asociación, etc.) son debidas a un único factor "inteligencia".

Los factores comunes, generalmente pocos, son los que explican todas las variables, y no se pueden observar o medir: son implícitos. El objetivo del análisis factorial consiste en la mayoría de los casos en averiguar si el modelo, con los factores –en número y significado– que hemos elaborado, se corresponde con la realidad observada a través de nuestros datos.

Para completar el modelo se añade un factor específico a cada variable, que recoge la variabilidad que no consigue explicar el conjunto de factores comunes; así, en un modelo con *n* variables observadas y dos factores comunes:

$$\begin{aligned} X_1 &= \lambda_{11} F_1 + \lambda_{12} F_2 + e_1 \\ X_2 &= \lambda_{21} F_1 + \lambda_{22} F_2 + e_2 \\ \ldots &\quad \ldots \quad \ldots \quad \ldots \\ X_n &= \lambda_{n1} F_1 + \lambda_{n2} F_2 + e_n \end{aligned}$$

Se denomina saturaciones a los coeficientes del modelo λ_{ij}, que deben ser estimados: λ_{ij} es la saturación de X_i en F_j.

Las variables deben ser cuantitativas y estar correlacionadas, ya que si no lo están no existiría información compartida que pudiera ser explicada mediante factores comunes. Para la estimación de los coeficientes se añaden otras condiciones o hipótesis instrumentales que no necesitan ser comprobadas: Todos los factores tienen media cero y correlación nula entre sí, y los factores comunes tienen varianza 1. Existen varios procedimientos para la estimación de los coeficientes, siendo uno de ellos el método de componentes principales; por esa razón a menudo se confunden ambas técnicas, que son conceptualmente distintas.

A partir del modelo anterior se puede expresar la varianza de cada una de las variables X_i del siguiente modo:

$$\sigma^2_i = \lambda^2_{i1} + \lambda^2_{i2} + \varphi_i \qquad (\varphi_i : \text{varianza de la componente específica})$$

Es decir, la varianza de cada una de las variables es igual a la suma de los cuadrados de las saturaciones, que se denomina comunalidad, más un componente residual o especificidad. Ambas cantidades se expresan a menudo en porcentaje o fracción de la varianza total de la variable. Cuanto más altas sean las comunalidades, mejor funciona el modelo factorial.

Cuando el número de factores comunes es mayor que 1 el problema de la estimación de los coeficientes del modelo o saturaciones no está determinado: existen infinitas soluciones. Encontrada una solución, se puede obtener una transformación suya –rotación– que también es una solución válida. Es posible por lo tanto buscar –dentro de las infinitas soluciones existentes– aquellas que permiten una mejor interpretación del modelo. Thurstone estableció una serie de criterios o principios de estructura simple para facilitar la búsqueda de las mejores soluciones: cada factor debe tener pocas saturaciones altas y el resto muy bajas (lo que facilita la interpretación de los factores en términos de las variables), cada variable debe estar muy saturada en un solo factor (lo que facilita la interpretación de las variables en términos de los factores) y, por último, los factores deben ser distintos (ya que se supone que explican aspectos diferentes).

De acuerdo con estos principios, se han desarrollado procedimientos para obtener rotaciones adecuadas:

Rotaciones ortogonales: Mantienen la independencia o incorrelación de los factores:

- Varimax: max. varianza(λ_{ij}^2) para cada F_i (simplifica los factores)
- Quartimax: max. varianza(λ_{ij}^4) para cada x_j (simplifica las variables)
- Equamax: intermedia (trata de simplificar factores y variables)

Rotaciones oblicuas (los factores estarán correlacionados entre sí):

- Oblimin
- Promax

En general será adecuada la rotación varimax, ya que tendremos muchas variables y pocos factores, y solo necesitamos simplificar la expresión de los factores. En un modelo con muchos factores puede ser conveniente utilizar quartimax, para simplificar la expresión de cada variable. En algunos casos una rotación oblicua puede ser más satisfactoria, mejorando la interpretabilidad aunque con el sacrificio de la pérdida de independencia de los factores.

La condición más importante que debemos comprobar para la correcta aplicación del Análisis Factorial es la de correlación entre las variables originales. Para ello se utilizan los mismos procedimientos descritos en relación con el Análisis de Componentes Principales: Matriz de correlaciones, matriz de valores *p* de los contrastes de nulidad, determinante de la matriz de correlaciones y test de esfericidad de Bartlett. En el análisis factorial se emplea habitualmente un indicador denominado *KMO* (por sus autores Kaiser-Meyer-Olkin) o índice de adecuación muestral:

$$KMO = \frac{\sum\sum r_{ij}^2}{\sum\sum r_{ij}^2 + \sum\sum \rho_{ij}^2}$$

Donde r_{ij}^2 es el cuadrado del coeficiente de correlación lineal entre las variables i y j, y ρ_{ij}^2 el cuadrado del coeficiente de correlación parcial entre dichas variables, excluyendo de ambas sumas los casos $i = j$. La correlación parcial no contribuye a la justificación de factores comunes que expliquen simultáneamente

todas las variables, por lo que, cuanto menor sea la suma de correlaciones parciales, en relación con las correlaciones no parciales o bivariantes, más adecuados serán los datos para un modelo factorial. *KMO* toma valores entre 0 y 1, y los propios autores establecieron la siguiente guía de interpretación:

- <0,5 inaceptable
- 0,5 - 0,6 malo
- 0,6 - 0,7 regular
- 0,7 - 0,8 bueno
- >0,8 excelente

Dado que a menudo existe una cierta confusión entre Análisis Factorial y Análisis de Componentes Principales, es conveniente resumir las principales diferencias entre ambos métodos:

- En el Análisis Factorial cada variable es una función de los factores, en el Análisis de Componentes Principales cada componente es función de las variables originales (los modelos son inversos)
- Análisis de Componentes Principales tiene una única solución, Análisis Factorial infinitas (rotaciones)
- En el Análisis Factorial existe un modelo previo.
- En el Análisis Factorial existen múltiples métodos de estimación de los coeficientes (Análisis de Componentes Principales es uno de ellos)
- El Análisis de Componentes Principales analiza la matriz de covarianzas o de correlaciones, el Análisis Factorial la matriz de correlaciones corregida, sustituyendo los unos de la diagonal por la comunalidad estimada.

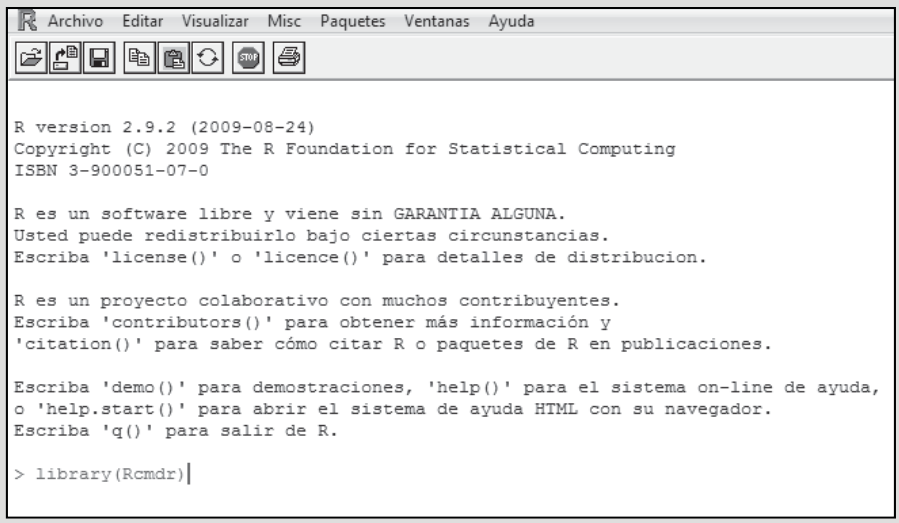
CUADRO XIII.2. Análisis factorial

Aplicación con R.

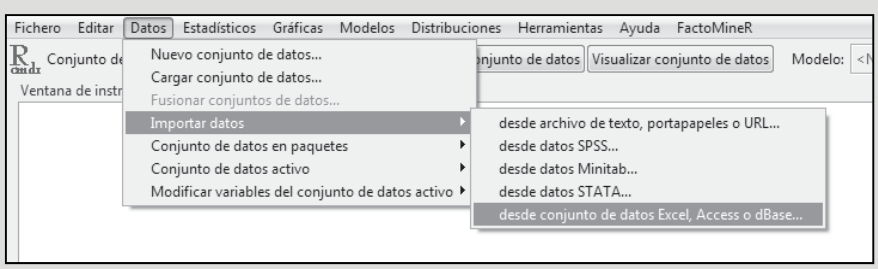
EJEMPLO. En el archivo **Cuadro III.2.xls** hay datos de variables limnológicas de varios lagos Neotropicales. Estas variables indican el grado de productividad del lago (concentración de nutrientes y carbono orgánico disuelto) y la adecuación del hábitat en lo que se refiere a sus condiciones para la vida (profundidad, pH, conductividad, oxígeno disuelto y temperatura). Estos dos factores, productividad y hábitat, podrían explicar razonablemente la correlación observada entre las distintas variables. El objetivo del análisis factorial en este contexto es comprobar si este modelo responde razonablemente a la realidad y es, por lo tanto, adecuado para explicar nuestros datos.

Paso 1. Iniciamos R y cargamos el paquete «Rcmdr» introduciendo la orden «library(Rcmdr)». Si el paquete no está instalado, la instalación se realiza como se explica en el Apéndice II.

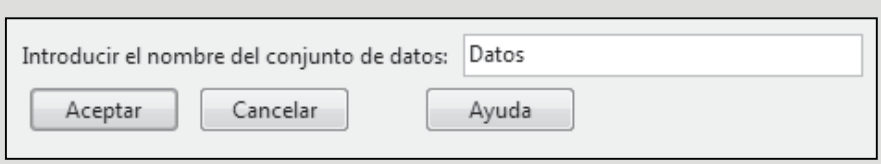
CUADRO XIII.2. Análisis factorial


Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> library(Rcmdr)

Paso 2. Después de haber seleccionado el directorio, como se explica en el Apéndice II, vamos a «Datos» ⇨ «Importar datos» ⇨ «desde conjunto de datos Excel, Access o dBase».


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda FactoMineR
Nuevo conjunto de datos...
Cargar conjunto de datos...
Fusionar conjuntos de datos...
Importar datos
Conjunto de datos en paquetes
Conjunto de datos activo
Modificar variables del conjunto de datos activo
desde archivo de texto, portapapeles o URL...
desde datos SPSS...
desde datos Minitab...
desde datos STATA...
desde conjunto de datos Excel, Access o dBase...
Visualizar conjunto de datos
Modelo:

Paso 3. Aparece un pequeño cuadro de diálogo para introducir el nombre del conjunto de datos, en el que dejamos el nombre por defecto «Datos» y pulsamos «Aceptar». Ahora aparecen los archivos en el directorio de trabajo, en el que elegimos **Cuadro XIII.2.xls** y pulsamos «Abrir».

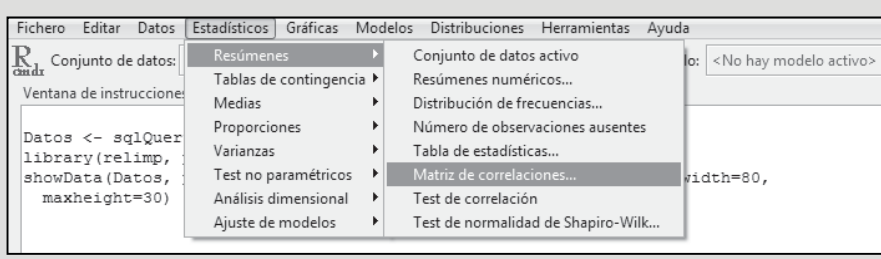


CUADRO XIII.2. *(Continuación)*[R]

Paso 4. En R Commander pulsamos en «Visualizar conjunto de datos». Es importante inspeccionar los datos y comprobar que no existan líneas, lo cual a veces ocurre al final del archivo, que tengan celdas con el código <NA>. Si existen estas líneas, hay que eliminarlas.

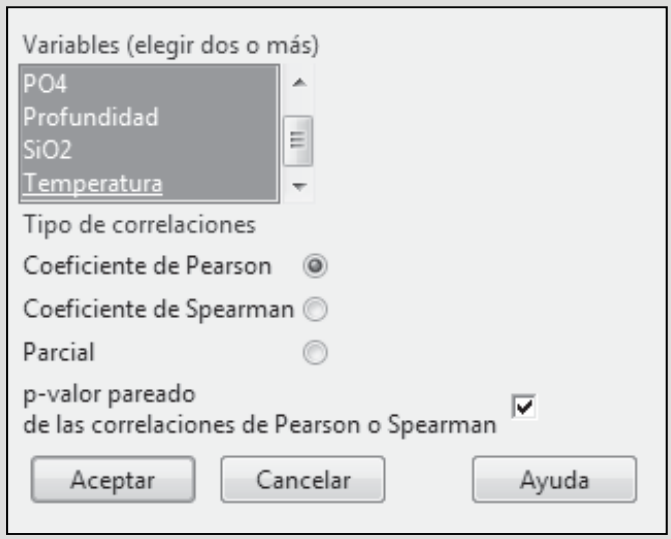
	Area	Lago	NO2	NO3	NH4	PO4	SiO2	Carbono	Profundidad
1	Amazonas	Correo	0.121333333	0.3600000	6.321000	1.08766667	205.58100	5.828500	8.700
2	Amazonas	Correo	0.097000000	0.3463333	4.940667	0.92033333	192.24933	5.336333	9.000
3	Amazonas	Correo	0.191000000	0.4020000	4.587000	1.02500000	198.49000	5.834000	6.200
4	Amazonas	Tarapoto	0.085500000	0.4295000	3.344000	0.47300000	219.94700	6.159500	6.100
5	Amazonas	Tarapoto	0.112666667	0.4040000	4.236333	0.71533333	223.95700	6.144000	13.500
6	Amazonas	Tarapoto	0.180500000	0.4230000	4.285500	0.39600000	225.56250	6.723000	9.200
7	Amazonas	Tarapoto	0.134666667	0.4700000	9.079333	0.64366667	198.08533	6.146000	6.500
8	Amazonas	Yahuarcaca	0.199000000	0.8150000	3.900500	0.54425000	305.18200	6.145000	7.250
9	Amazonas	Yahuarcaca	0.140500000	0.5777500	3.132500	0.30125000	306.13825	6.489750	6.900
10	Amazonas	Yahuarcaca	0.226000000	0.7710000	4.246667	0.47233333	429.76333	3.222333	6.400
11	Amazonas	Yahuarcaca	0.268333333	0.6203333	4.232000	0.38500000	325.05900	5.209333	6.300
12	Andes	Fuquene	0.038000000	0.2920000	3.657000	0.01650000	22.91750	3.346000	1.650
13	Andes	Fuquene	0.043500000	1.3355000	2.772500	0.01750000	22.45850	3.714500	2.200
14	Andes	Fuquene	0.048000000	1.2635000	2.354000	0.00000000	25.62700	3.896500	4.300
15	Andes	Fuquene	0.019500000	0.8650000	2.647000	0.00000000	19.90450	3.527000	0.925
16	Andes	Iguaque	0.052000000	0.1825000	0.993500	0.00000000	0.12750	2.800000	4.500
17	Andes	Iguaque	0.033500000	0.4565000	0.816000	0.00000000	0.08000	2.700000	4.600
18	Andes	Guatavita	0.010000000	0.5950000	7.957000	0.00000000	45.49967	3.209667	9.600
19	Andes	Guatavita	0.006666667	0.4696667	6.521000	0.00000000	22.43600	1.855000	19.750
20	Andes	Tota	0.069333333	0.8716667	2.317333	0.02366667	18.25000	1.585000	50.000
21	Andes	Tota	0.077333333	0.9320000	1.749333	0.00000000	19.76633	1.602500	46.000
22	Andes	Tota	0.085000000	0.9900000	4.440000	0.00000000	19.69133	2.257500	28.100
23	Andes	Tota	0.043000000	1.4380000	1.216000	0.00000000	21.43533	1.700000	27.700
24	Andes	Tota	0.052666667	1.5283333	1.776000	0.01533333	22.00467	1.970667	23.500
25	Caribe	Momil	0.100000000	1.1687500	6.054500	1.99425000	184.21975	4.821000	1.700
26	Caribe	S. Sebastian	0.118500000	1.9260000	6.391500	1.84625000	176.99975	4.700000	2.600
27	Caribe	S. Sebastian	0.181500000	1.4212500	3.051000	0.24075000	177.73450	5.507000	2.750
28	Caribe	Purisima	0.086000000	0.7687500	3.283000	0.26150000	170.24425	4.871667	2.000
29	Caribe	Purisima	0.082000000	0.4740000	3.252000	0.09000000	168.30400	4.554000	2.320

Paso 5. Antes de aplicar el Análisis Factorial debemos comprobar si nuestras variables están correlacionadas. Pulsamos en el menú principal «Estadísticos» ⇨ «Resúmenes» ⇨ «Matriz de correlaciones».



Paso 6. En el cuadro de diálogo que aparece seleccionamos todas las variables, y marcamos «Coeficiente de Pearson» y «p-valor pareado», y pulsamos «Aceptar».

CUADRO XIII.2. *(Continuación)*[R]



Paso 7. En la ventana de resultados de R Commander aparece la matriz de correlaciones, la de valores *p* de los contrastes de nulidad, y otra de valores *p* corregidos para múltiples contrastes. Es más correcto utilizar esta última cuando exploramos la existencia de posibles correlaciones nulas, y la primera cuando solo nos interesa contrastar un único coeficiente de correlación concreto. Como podemos observar en la matriz de valores *p* corregidos «Adjusted p-values», hay bastantes coeficientes que pueden ser nulos ($p > 0{,}05$). Los correspondientes coeficientes de correlación, lógicamente son pequeños. En general, cuanto más altos sean los coeficientes (y más pequeños, próximos a cero, los valores *p*), mejor funcionará el Análisis Factorial. En este ejemplo, la matriz de valores *p* nos dice claramente que algunas variables no están correlacionadas con las restantes y, por lo tanto, no se adecúan a un modelo factorial, como veremos posteriormente.

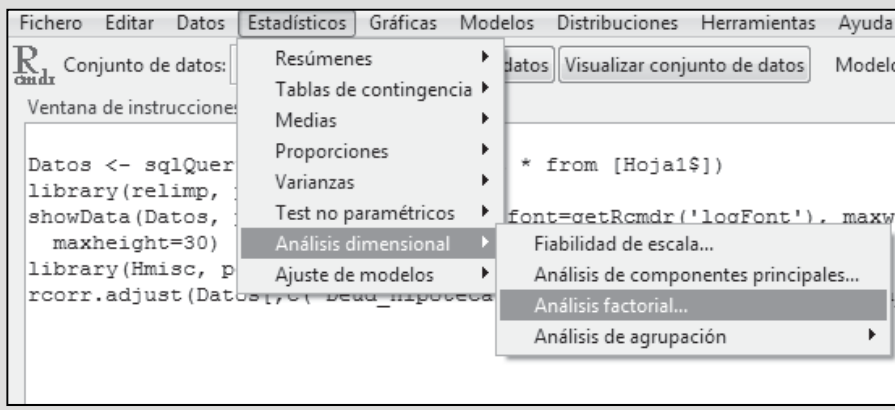
	Carbono	Conductividad	NH4	NO2	NO3	Oxigeno	pH	PO4	Profundidad	SiO2	Temperatura
Carbono	1.00	0.22	0.39	0.61	-0.28	-0.64	-0.22	0.52	-0.59	0.73	0.82
Conductividad	0.22	1.00	0.02	0.58	0.35	-0.49	0.12	0.27	-0.28	0.60	0.59
NH4	0.39	0.02	1.00	0.19	-0.15	-0.46	-0.37	0.53	-0.23	0.34	0.41
NO2	0.61	0.58	0.19	1.00	-0.07	-0.67	-0.16	0.40	-0.19	0.86	0.72
NO3	-0.28	0.35	-0.15	-0.07	1.00	-0.15	0.21	0.14	0.15	-0.19	-0.06
Oxigeno	-0.64	-0.49	-0.46	-0.67	-0.15	1.00	0.23	-0.61	0.42	-0.73	-0.68
pH	-0.22	0.12	-0.37	-0.16	0.21	0.23	1.00	-0.40	0.17	-0.12	-0.17
PO4	0.52	0.27	0.53	0.40	0.14	-0.61	-0.40	1.00	-0.31	0.50	0.64
Profundidad	-0.59	-0.28	-0.23	-0.19	0.15	0.42	0.17	-0.31	1.00	-0.37	-0.45
SiO2	0.73	0.60	0.34	0.86	-0.19	-0.73	-0.12	0.50	-0.37	1.00	0.88
Temperatura	0.82	0.59	0.41	0.72	-0.06	-0.68	-0.17	0.64	-0.45	0.88	1.00

Adjusted p-values (Holm's method)

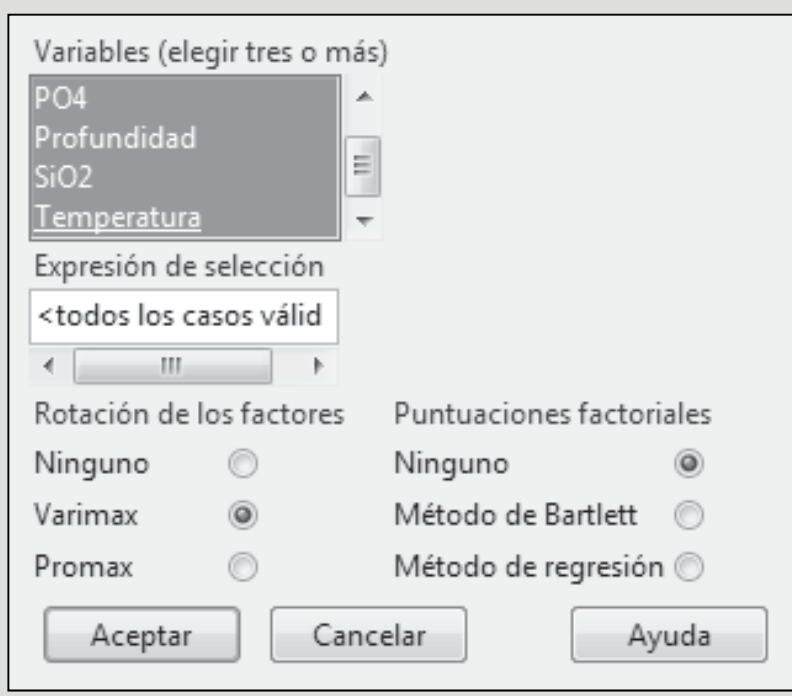
	Carbono	Conductividad	NH4	NO2	NO3	Oxigeno	pH	PO4	Profundidad	SiO2	Temperatura
Carbono		1.0000	1.0000	0.0216	1.0000	0.0085	1.0000	0.1350	0.0287	0.0004	0.0000
Conductividad	1.0000		1.0000	0.0418	1.0000	0.2602	1.0000	1.0000	1.0000	0.0257	0.0287
NH4	1.0000	1.0000		1.0000	1.0000	0.4464	1.0000	0.1296	1.0000	1.0000	0.8534
NO2	0.0216	0.0418	1.0000		1.0000	0.0031	1.0000	0.9733	1.0000	0.0000	0.0005
NO3	1.0000	1.0000	1.0000	1.0000		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Oxigeno	0.0085	0.2602	0.4464	0.0031	1.0000		1.0000	0.0207	0.8051	0.0003	0.0029
pH	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000		0.9138	1.0000	1.0000	1.0000
PO4	0.1350	1.0000	0.1296	0.9733	1.0000	0.0207	0.9138		1.0000	0.2222	0.0082
Profundidad	0.0287	1.0000	1.0000	1.0000	1.0000	0.8051	1.0000	1.0000		1.0000	0.4670
SiO2	0.0004	0.0257	1.0000	0.0000	1.0000	0.0003	1.0000	0.2222	1.0000		0.0000
Temperatura	0.0000	0.0287	0.8534	0.0005	1.0000	0.0029	1.0000	0.0082	0.4670	0.0000	

CUADRO XIII.2. *(Continuación)*[R]

Paso 8. Para aplicar el Análisis Factorial elegimos en el menú principal de R Commander las opciones «Estadísticos» ⇨ «Análisis dimensional» ⇨ «Análisis factorial…».

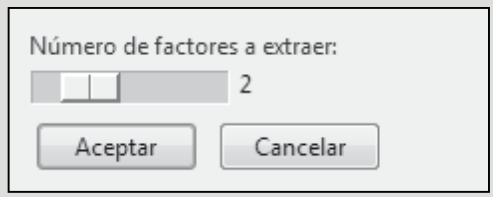


Paso 9. En el cuadro de diálogo que aparece seleccionamos las variables que queremos utilizar (en este caso todas), marcamos «Varimax» en «Rotación de los factores» y «Ninguno» en «Puntuaciones factoriales».



CUADRO XIII.2. *(Continuación)*[R]

Paso 10. En el cuadro «Número de componentes a retener», movemos el indicador a 2.



Paso 11. En la ventana de resultados de R commander aparecen en primer lugar las especificidades «Uniquenesses» de cada variable, expresadas en fracción de varianza. Cuatro variables tienen especificidad baja: «Conductividad» (0,005), «Temperatura» (0,084), «Carbono» (0,15) y «SiO2» (0,155), por lo tanto, la comunalidad (el resto hasta uno) es alta. La mayor parte de la variabilidad está explicada por los factores comunes y estas cuatro variables responden satisfactoriamente a nuestro modelo. Las otras variables, por el contrario, tienen especificidades muy elevadas, y no son explicadas adecuadamente por este modelo.

```
Uniquenesses:
      Carbono Conductividad            NH4            NO2            NO3
        0.150          0.005          0.744          0.359          0.724

 Oxigeno            pH           PO4   Profundidad          SiO2   Temperatura
   0.419         0.876         0.594         0.742         0.155         0.084
```

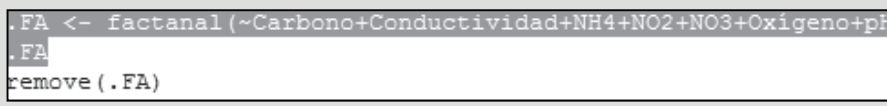
Paso 12. Los resultados muestran un test para probar la hipótesis de que dos factores son suficientes. Este test está basado en la reconstrucción de la matriz de correlaciones a partir del modelo factorial, y en las diferencias entre esa matriz y la matriz de correlaciones original. Si el modelo funciona bien, las diferencias entre ambas matrices de correlaciones deben ser prácticamente nulas. En nuestra aplicación el valor *p* del contraste vale 0,00168, lo que nos obliga a rechazar la hipótesis de que dos factores son suficientes, es decir, se requiere un modelo más completo. Esta conclusión es consistente con la observación anterior de que muchas variables no están suficientemente explicadas por los dos factores ya que tienen una comunalidad muy baja. Debe señalarse que este test es muy sensible, y cuando la muestra es grande tiende a rechazar la hipótesis de suficiencia del modelo incluso cuando el análisis factorial funciona muy bien, por lo que debe interpretarse con prudencia. El valor *p* muy pequeño no significa necesariamente que el modelo es inadecuado.

```
Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 63.28 on 34 degrees of freedom.
The p-value is 0.00168
```

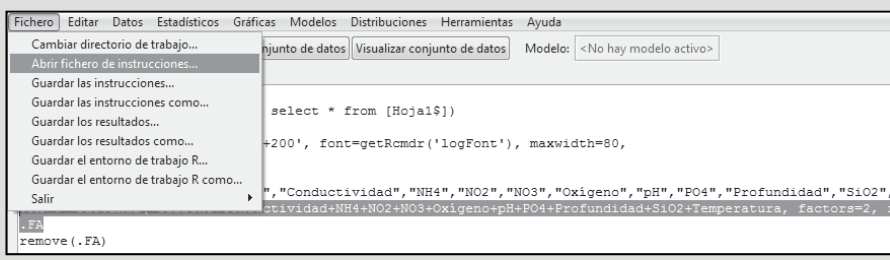
CUADRO XIII.2. *(Continuación)*[R]

Paso 13. Para facilitar la interpretación de las saturaciones («Loadings»), obtener el valor de KMO y representar los datos, se copian las instrucciones del Análisis Factorial, que se muestran a continuación y en la siguiente ventana (no se muestra la instrucción completa).

```
.FA <- factanal(~Carbono+Conductividad+NH4+NO2+NO3+Oxígeno+pH
+PO4+Profundidad+SiO2+Temperatura, factors=2, rotation="varimax",
scores="none", data=Datos)
.FA
```



Paso 14. A continuación se carga el archivo de instrucciones **Cuadro XIII.2.R** seleccionando «Fichero» ⇨ «Abrir fichero de instrucciones…» y no se guarda la salida.



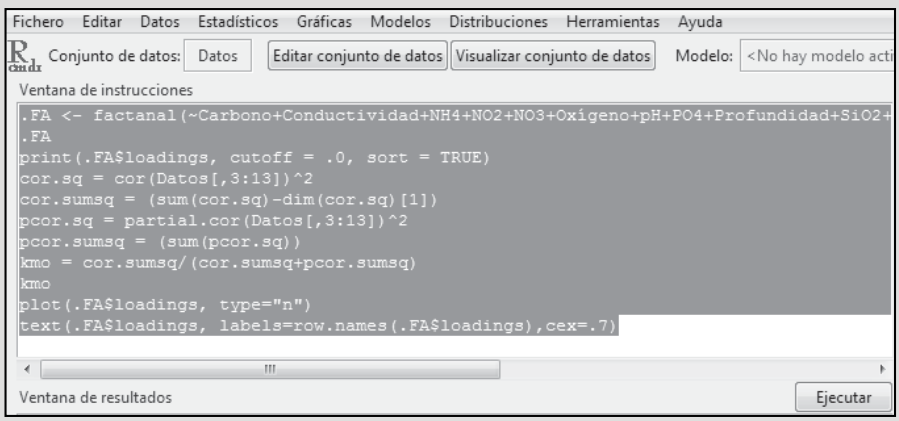
Paso 15. Se pegan las instrucciones que se copiaron en el paso 13, arriba de las instrucciones que se importaron del archivo **Cuadro XIII.2.R**. Las instrucciones del archivo **Cuadro XIII.2.R**, que se muestran a continuación, permiten calcular el índice KMO.

```
cor.sq = cor(Datos[,3:13])^2
cor.sumsq = (sum(cor.sq)-dim(cor.sq)[1])
pcor.sq = partial.cor(Datos[,3:13])^2
pcor.sumsq = (sum(pcor.sq))
kmo = cor.sumsq/(cor.sumsq+pcor.sumsq)
kmo
```

Las variables implicadas en este ejemplo están en las columnas 3 a 13 del conjunto de datos (para otra aplicación deben modificarse estos números las dos veces que aparecen en las instrucciones), y el conjunto de datos activo es «Datos» (el nombre por defecto que asigna Rcmdr). Si las variables a utilizar no son contiguas, en lugar de la referencia 3:13 (variables desde la 3 a la 13) debemos utilizar los números de columna separados por comas, (por ejemplo Datos[,1,3,5,6]. Estas instrucciones para el cálculo de KMO solo funcionarán correctamente si nuestros datos no contienen valores perdidos o no definidos.

CUADRO XIII.2. *(Continuación)*[R]

Por último, se selecciona todo y se pulsa «Ejecutar»



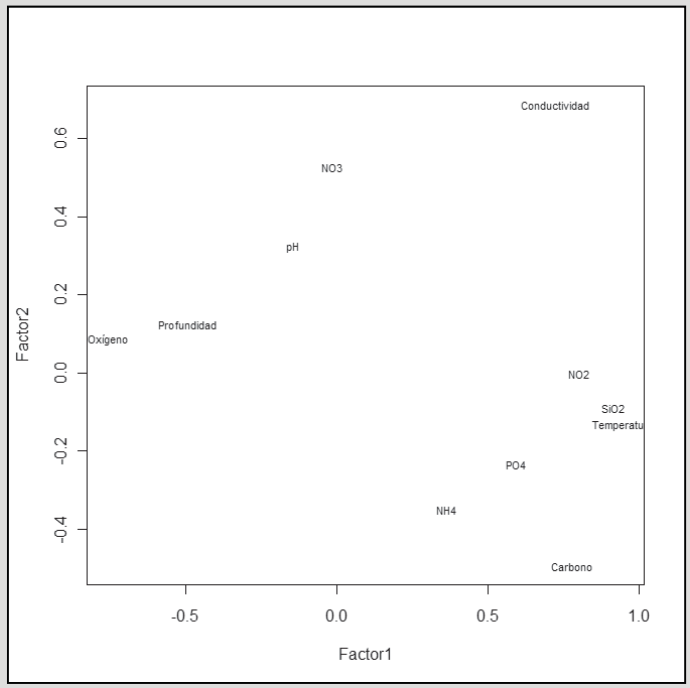
Paso 16. En la ventana de resultados vemos de nuevo las saturaciones («Loadings»), pero ahora ordenadas por factores.

```
Loadings:
               Factor1 Factor2
Carbono          0.778  -0.494
Conductividad    0.724   0.686
NO2              0.801  -0.003
Oxígeno         -0.757   0.087
PO4              0.592  -0.235
SiO2             0.915  -0.091
Temperatura      0.947  -0.135
NO3             -0.014   0.526
NH4              0.364  -0.351
pH              -0.144   0.322
Profundidad     -0.492   0.123
```

El primer factor está asociado a la cantidad de carbono orgánico disuelto, el NO_2, SiO_4, es decir, nutrientes y, por tanto, indicadores de la capacidad de producción del lago. Sin embargo, también está asociado a indicadores de las características físicas y químicas del lago, como la concentración de oxígeno y la temperatura. Por lo tanto, el primer eje separa lagos con bajas temperaturas, alta cantidad de oxígeno y baja capacidad de producción (izquierda del primer eje); de los lagos con mayor temperatura, más productivos y menos concentración de oxígeno. El segundo factor está asociado preferentemente a la conductividad, pero veremos más adelante que este factor absorbe un porcentaje muy pequeño de la variabilidad explicada.

En el gráfico siguiente se pueden ver las saturaciones mostradas en la tabla anterior, donde se observa el patrón explicado anteriormente.

CUADRO XIII.2. *(Continuación)*[R]



Otro de los resultados mostrados es la variabilidad explicada por los factores: el primero explica un 43,9% y el segundo el 12,0%. Los dos primeros factores en su conjunto explican el 55,9% de la variabilidad total, un porcentaje bajo.

	Factor1	Factor2
SS loadings	4.824	1.323
Proportion Var	0.439	0.120
Cumulative Var	0.439	0.559

En la ventana de resultados aparece KMO = 0,656 y, como se mencionó en la introducción del Análisis Factorial, ese valor está dentro del rango de regular, por lo que la adecuación de los datos de la muestra a un modelo factorial no es muy buena.

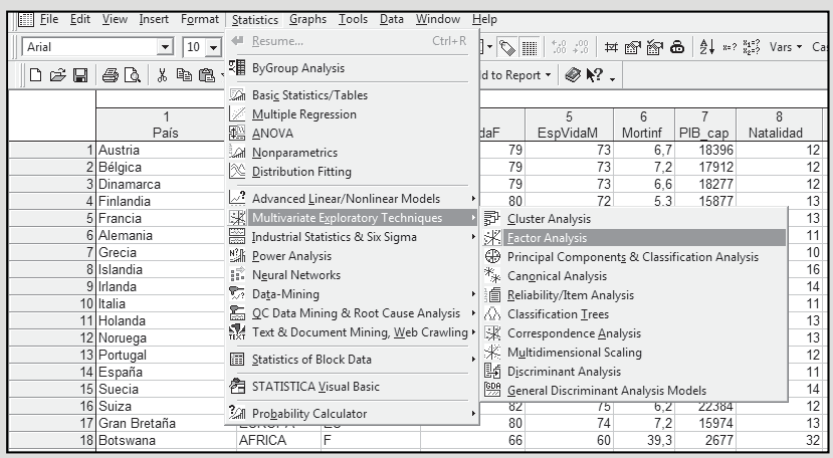
En resumen, el modelo parece adecuado únicamente para algunas de las variables objeto de estudio: conductividad, temperatura, oxígeno, carbono disuelto, NO_2 y SiO_2. El resto de las variables no responden a un modelo factorial, por lo que éste debe ser reformulado. El hecho de que todas las variables con comunalidad elevada estén saturadas en el primer factor orienta hacia un modelo con un único factor, relacionado con la productividad y con las características del hábitat, en lo que se refiere a su temperatura, concentración de oxígeno y conductividad. Las variables excluidas, aquellas no explicadas razonablemente por el modelo, no parecen estar relacionadas con el primer factor y, por lo tanto, carecen de interés en este contexto.

CUADRO XIII.2. *(Continuación)*[STATISTICA]

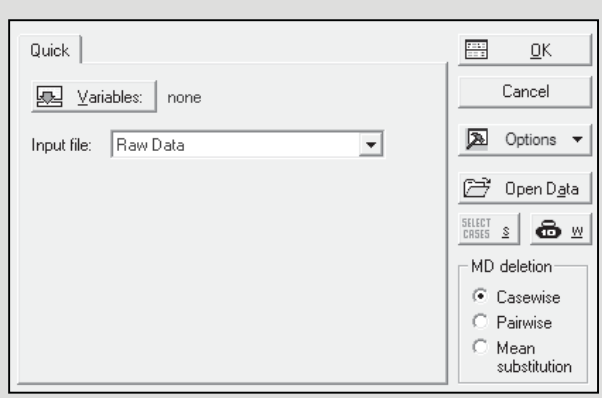
APLICACIÓN CON STATISTICA.

EJEMPLO. Utilizaremos los datos contenidos en el archivo **Cuadro XIII.2.sta**. Las variables indican, para diferentes países, la esperanza de vida masculina y femenina al nacer (en años de vida), las tasas de mortalidad, mortalidad infantil, natalidad, y fertilidad, el Producto Interior Bruto per cápita (en miles de dólares anuales), y la tasa de alfabetización de hombres y mujeres (en porcentaje). Todas estas variables están relacionadas con el grado de desarrollo de cada país, aunque algunas de ellas parecen tener un contenido claramente económico y otras quizás un contenido cultural o de estilo de vida, por lo que supondremos que todas ellas son el resultado de un factor que llamaremos «Económico» y otro «Cultural» independiente del anterior.

Paso 1. Una vez abierto el archivo de datos con STATISTICA, para aplicar el Análisis Factorial elegimos en el menú principal las opciones «Statistics» ⇨ «Multivariate Exploratory Techniques» ⇨ «Factor Analysis».

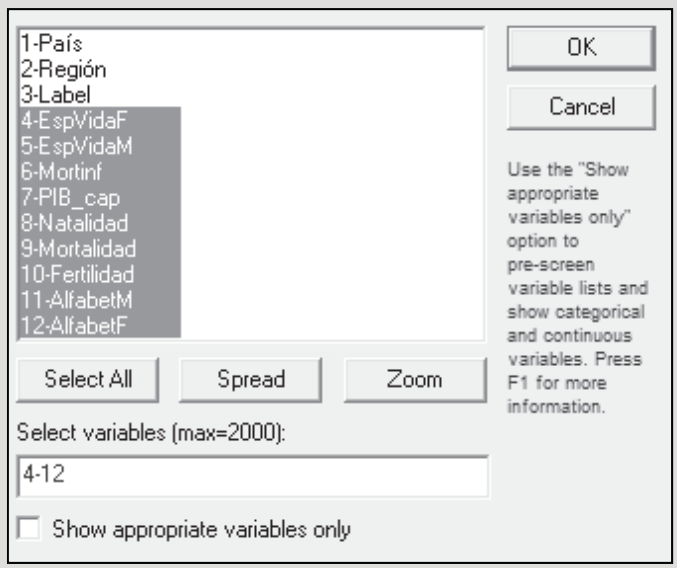


Paso 2. En el cuadro de diálogo que aparece pulsamos «Variables».

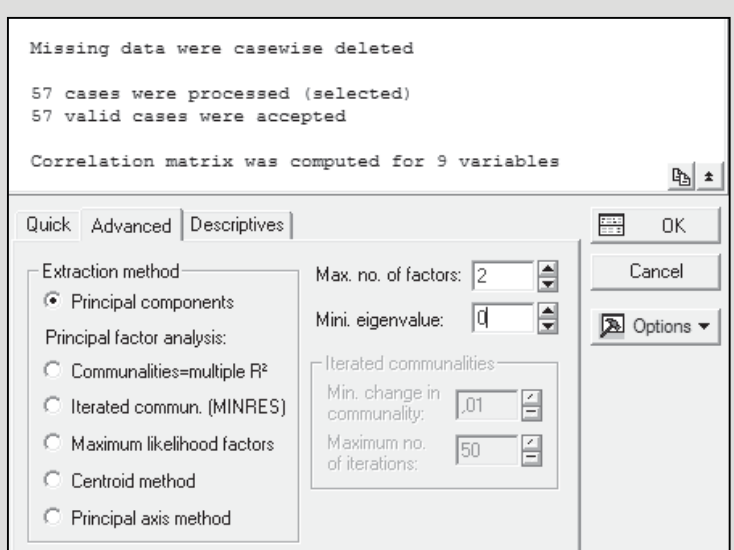


CUADRO XIII.2. *(Continuación)*[STATISTICA]

Paso 3. Seleccionamos todas las variables cuantitativas de nuestros datos, desde la 4 hasta la 12.

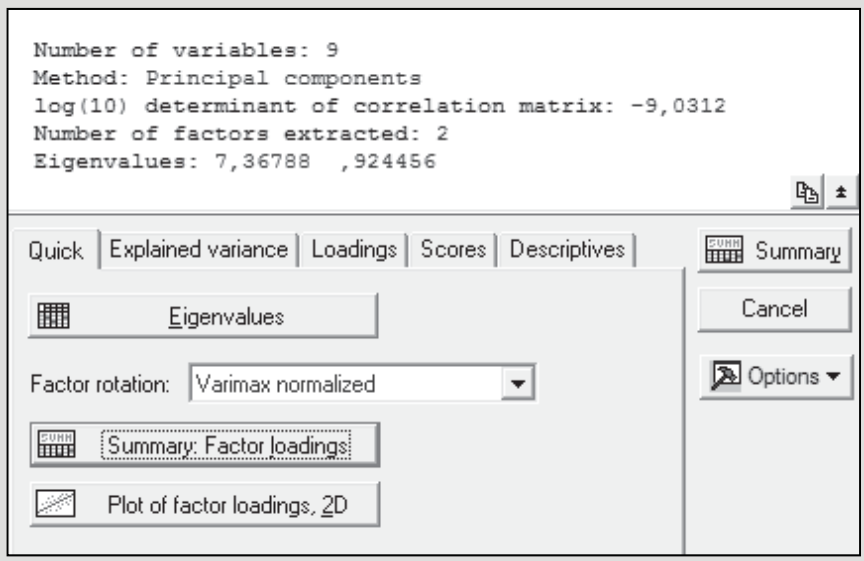


Paso 4. En la siguiente ventana, en la pestaña «Advanced» ponemos 2 factores en «Max. no. of factors:» y 0 en «Mini. eigenvalue:» (si dejamos el valor por defecto 1, seleccionará el número de factores con autovalores mayores que 1 hasta un máximo de 2). En «Extraction method» seleccionamos «Principal components» que no exige ningún requisito de las variables. Si la mayoría de las variables tuviesen una distribución Normal, una buena opción podría ser «Maximum likelihood factors».



CUADRO XIII.2. *(Continuación)*[STATISTICA]

Paso 5. En la ventana que aparece pulsamos en primer lugar Quick para elegir el método de rotación: «Varimax normalized».

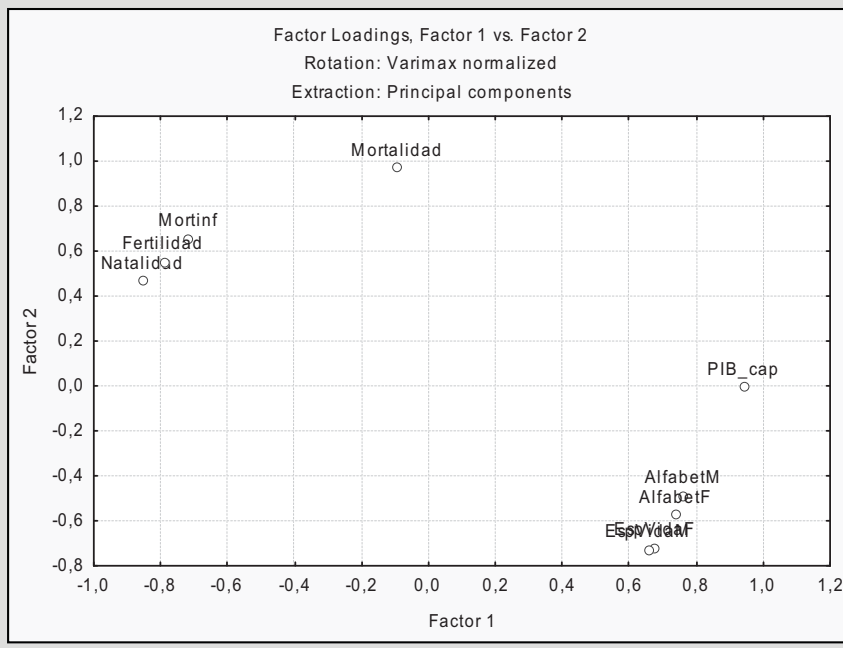


Paso 6. Al pulsar en la ventana anterior en «Summary: Factor loadings»; obtenemos los siguientes resultados. El primer factor está más relacionado con Mortalidad infantil, PIB per cápita, tasas de natalidad y fertilidad, y tasas de alfabetización. El segundo está asociado a las variables esperanza de vida masculina y femenina, y tasa de mortalidad. El primer factor se corresponde con el que hemos denominado a priori «Económico», pero el segundo quizás podría denominarse «Supervivencia» o directamente «Mortalidad», ya que otras variables de contenido cultural parecen más relacionadas con el primer factor.

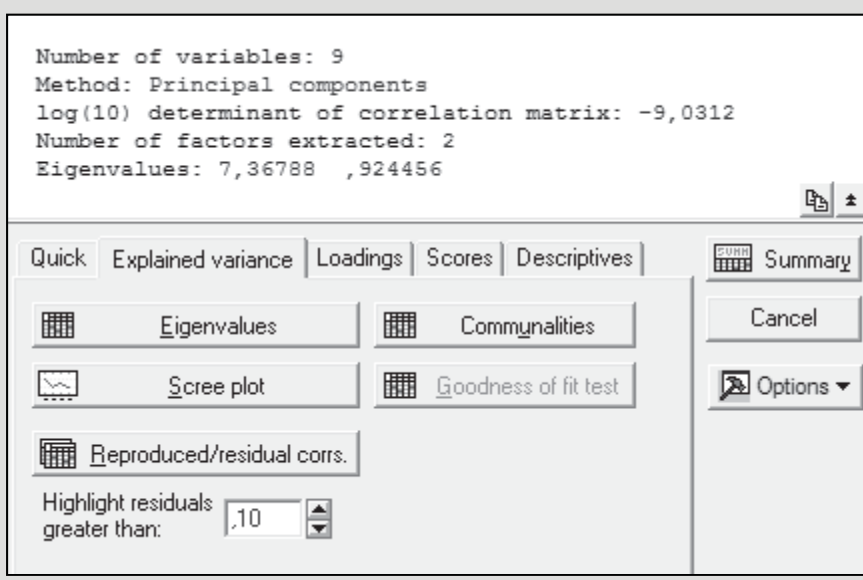
	Factor Loadings (Varimax normalized) (Cuadro XIII.2.sta) Extraction: Principal components (Marked loadings are >,700000)	
Variable	Factor 1	Factor 2
EspVidaF	0,674786	-0,717201
EspVidaM	0,655112	-0,726385
Mortinf	-0,718235	0,659323
PIB_cap	0,940151	-0,002465
Natalidad	-0,856391	0,468450
Mortalidad	-0,096097	0,975393
Fertilidad	-0,791497	0,554742
AlfabetM	0,761112	-0,485494
AlfabetF	0,737974	-0,569271
Expl.Var	4,777257	3,515075
Prp.Totl	0,530806	0,390564

CUADRO XIII.2. *(Continuación)*[STATISTICA]

Paso 7. En el paso 5 pulsamos en «Plot of factor loadings, 2D» y obtenemos el gráfico de saturaciones, que nos ayuda a interpretar los factores en términos de las saturaciones.



Paso 8. En la pestaña «Explained variance» pulsamos en «Communalities».



CUADRO XIII.2. *(Continuación)*[STATISTICA]

Paso 9. Las comunalidades indican la parte de varianza de cada variable explicada por el modelo factorial. Statistica nos presenta la comunalidad para el modelo con un solo factor, y con dos factores, además de los coeficientes de correlación múltiple, que generalmente se utilizan como estimación inicial de las comunalidades. El modelo de dos factores, el que hemos elegido, tiene altas comunalidades para todas las variables, por lo que todas ellas están explicadas satisfactoriamente por el modelo.

	Communalities (Cuadro XIII.2.sta) Extraction: Principal components Rotation: Varimax normalized		
Variable	From 1 Factor	From 2 Factors	Multiple R-Square
EspVidaF	0,455337	0,969714	0,993334
EspVidaM	0,429171	0,956807	0,988278
Mortinf	0,515861	0,950567	0,945073
PIB_cap	0,883883	0,883889	0,822898
Natalidad	0,733405	0,952851	0,980601
Mortalidad	0,009235	0,960626	0,894493
Fertilidad	0,626467	0,934206	0,972619
AlfabetM	0,579292	0,814996	0,962680
AlfabetF	0,544606	0,868676	0,972525

Paso 10. Pulsando en «Eigenvalues» podemos ver, por último, la importancia de cada factor. El primero explica el 81,86% y el segundo el 10,27%, por tanto entre ambos el 92,1%.

	Eigenvalues (Cuadro XIII.2.sta) Extraction: Principal components			
Value	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	7,367877	81,86530	7,367877	81,86530
2	0,924456	10,27173	8,292333	92,13703

En resumen, el modelo factorial con dos factores parece explicar razonablemente la variabilidad y las relaciones entre las variables de nuestro conjunto de datos, aunque quizás debamos reconsiderar la interpretación inicial de los dos factores a la luz del análisis de las saturaciones. Un análisis más profundo incluyendo variables adicionales podría aumentar nuestro conocimiento del problema y mejorar el modelo.

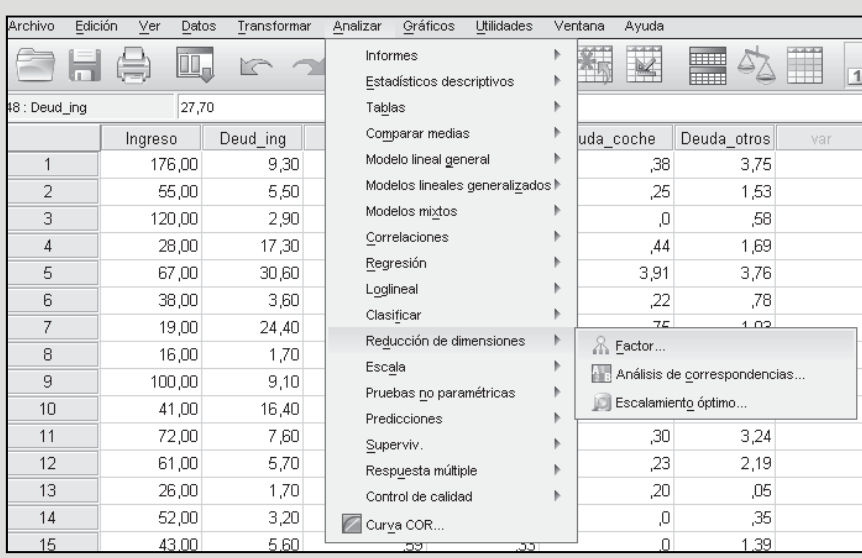
CUADRO XIII.2. *(Continuación)*[SPSS]

Aplicación con SPSS.

EJEMPLO. Tenemos datos de un conjunto de clientes de un banco en el archivo **Cuadro III.2.sav**. Las seis variables indican el nivel de ingresos, la relación (en porcentaje) entre deudas e ingresos, deudas con tarjeta de crédito, deudas por hipoteca, deudas por coche, y otras deudas. Supondremos que todas las variables de nuestro conjunto de datos, que están relacionadas con el gasto y el endeudamiento, dependen de dos factores; 1) Capacidad de gasto (relacionado con nivel de ingresos, patrimonio, expectativas de futuro, disponibilidad de crédito, etc.) y 2) Sostenibilidad (prudencia, ajuste gasto/ingreso, solvencia, hábitos de ahorro, etc.). Todos los clientes que forman nuestra muestra tienen un factor "capacidad de gasto", mayor o menor, difícil de cuantificar o de medir pero real, que incide en todas las variables financieras que sí podemos medir, y también tiene cada uno de ellos un factor "sostenibilidad" más alto o más bajo, relacionado con sus hábitos de gasto y con su forma de tomar decisiones en el ámbito financiero (más o menos prudente o equilibrada).

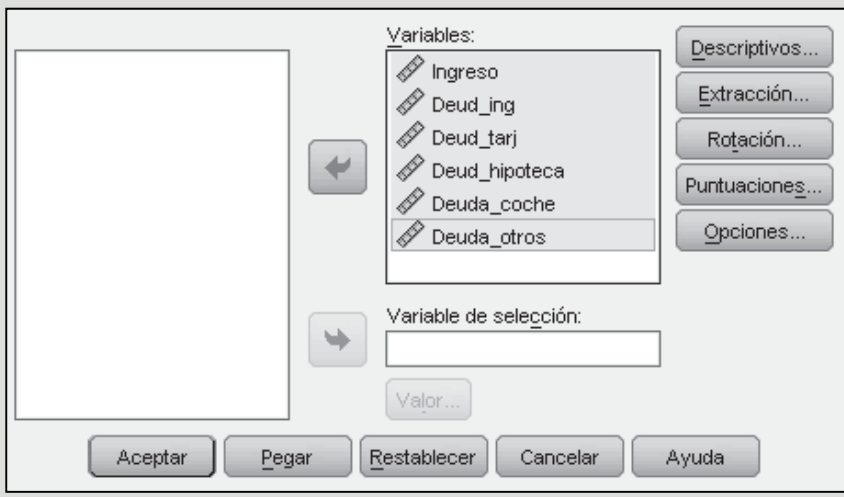
Podemos esperar que los clientes con mayor capacidad de gasto tengan más deudas de todo tipo y viceversa, y también que los clientes más prudentes tengan menos deudas. Estos dos factores podrían explicar razonablemente la correlación observada entre las distintas variables. El objetivo del análisis factorial en este contexto es comprobar si este modelo responde razonablemente a la realidad y es, por lo tanto, adecuado para explicar nuestros datos.

Paso 1. Se selecciona «Analizar» ⇨ «Reducción de dimensiones» ⇨ «Factor…».

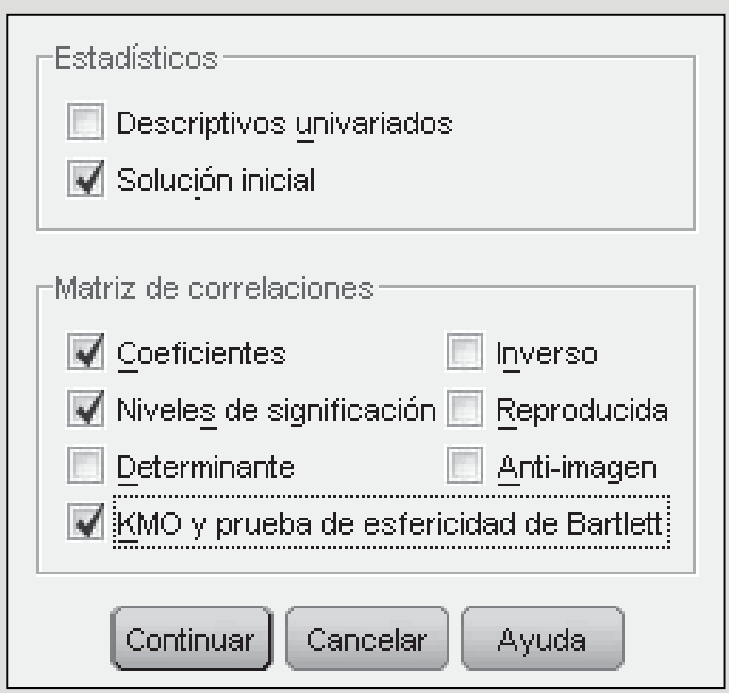


CUADRO XIII.2. *(Continuación)*[SPSS]

Paso 2. Se seleccionan todas las variables de interés y se pulsa en «Descriptivos…».

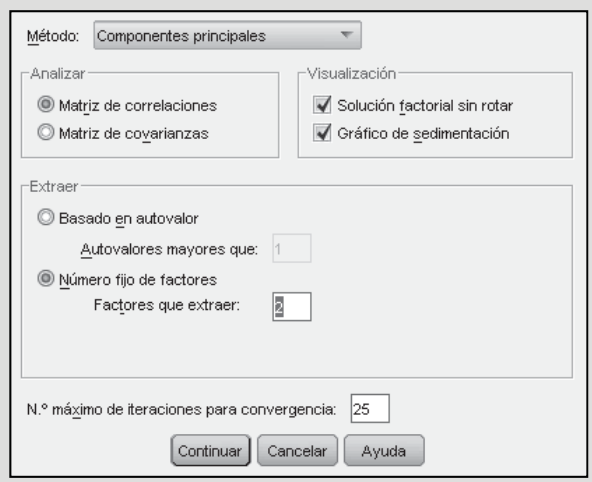


Paso 3. En la ventana que aparece se selecciona «Solución inicial», «Coeficientes», «Nivel de significación» y «KMO y prueba de esfericidad de Bartlett».

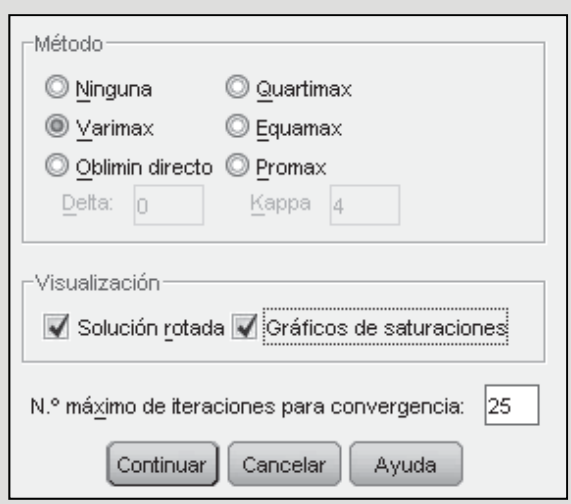


CUADRO XIII.2. *(Continuación)*[SPSS]

Paso 4. En la ventana del paso 2 se pulsa en «Extracción…». En el apartado «Método» mantenemos la opción por defecto, «Componentes principales», que no exige ningún requisito de las variables. Si la mayoría de las variables tuviesen una distribución Normal, otra opción podría ser «Maximum likelihood». En «Analizar» seleccionamos «Matriz de correlaciones» -ya que nuestras variables no son homogéneas, es decir, no están expresadas en las mismas unidades-. Es posible que con la opción «Matriz de covarianzas» se consiga explicar un mayor porcentaje de la varianza con los primeros ejes, pero esto sería incorrecto, ya que no se debe usar si no son homogéneas las variables. En «Visualización» dejamos la opción «Solución factorial sin rotar» y marcamos «Gráfico de sedimentación». Por último, en «Extraer» marcamos «Número fijo de factores» y escribimos el número que nos interesa: 2 factores.

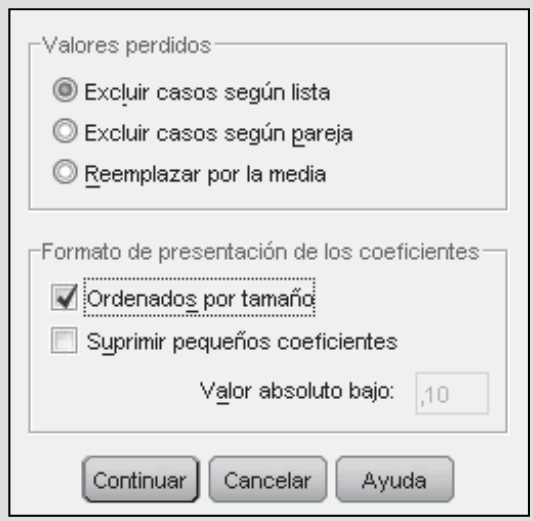


Paso 5. En la ventana del paso 2 se pulsa en «Rotación…». «En Método» se selecciona «Varimax» y marcamos «Gráfico de saturaciones».



CUADRO XIII.2. *(Continuación)*[SPSS]

Paso 6. En el cuadro de diálogo de «Opciones…» mantenemos la opción por defecto «Excluir casos según lista», ya que no tenemos valores perdidos. Marcamos «Ordenados por tamaño», para facilitar la interpretación de los factores.



Paso 7. En la tabla de correlaciones, la mitad superior es la matriz de correlaciones y la mitad inferior es la matriz de valores *p* de los contrastes de nulidad para cada uno de los coeficientes. En general es más sencillo observar esta última, ya que los valores *p* deben ser mayoritariamente menores que 0,05. Solo la variable ingreso muestra valores *p* grandes –y consecuentemente coeficientes de correlación pequeños– con las restantes. Por tanto, los ingresos no parecen muy correlacionados con las deudas por hipoteca y coche. En general se observa un nivel de correlación aceptable.

Matriz de correlaciones

		Ingreso	Deud_ing	Deud_tarj	Deud_hipoteca	Deuda_coche	Deuda_otros
Correlación	Ingreso	1,000	-,204	,537	,087	,063	,560
	Deud_ing	-,204	1,000	,437	,747	,730	,330
	Deud_tarj	,537	,437	1,000	,454	,429	,507
	Deud_hipoteca	,087	,747	,454	1,000	,988	,503
	Deuda_coche	,063	,730	,429	,988	1,000	,477
	Deuda_otros	,560	,330	,507	,503	,477	1,000
Sig. (Unilateral)	Ingreso		,021	,000	,194	,268	,000
	Deud_ing	,021		,000	,000	,000	,000
	Deud_tarj	,000	,000		,000	,000	,000
	Deud_hipoteca	,194	,000	,000		,000	,000
	Deuda_coche	,268	,000	,000	,000		,000
	Deuda_otros	,000	,000	,000	,000	,000	

CUADRO XIII.2. *(Continuación)*[SPSS]

Paso 8. El test de esfericidad de Bartlett (p < 0,001) indica que existe correlación. El índice KMO = 0,624 muestra un valor regular.

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,624
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	629,694
	gl	15
	Sig.	,000

Paso 9. Lo siguiente que podemos observar en los resultados es la tabla de comunalidades, que indican la fracción de varianza de cada variable que es explicada por los factores comunes. Vemos que son elevadas, por lo que todas las variables son explicadas en buena medida por el modelo factorial.

Comunalidades

	Inicial	Extracción
Ingreso	1,000	,912
Deud_ing	1,000	,827
Deud_tarj	1,000	,673
Deud_hipoteca	1,000	,925
Deuda_coche	1,000	,907
Deuda_otros	1,000	,714

Método de extracción: Análisis de Componentes principales.

Paso 10. El porcentaje de varianza explicado por cada uno de los dos factores se muestra en la tabla siguiente. El primer factor explica el 56,09%, el segundo el 26,51% y en conjunto los dos factores el 82,61%.

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3,366	56,096	56,096	3,366	56,096	56,096
2	1,591	26,515	82,611	1,591	26,515	82,611
3	,540	8,996	91,608			
4	,362	6,038	97,645			
5	,131	2,180	99,826			
6	,010	,174	100,000			

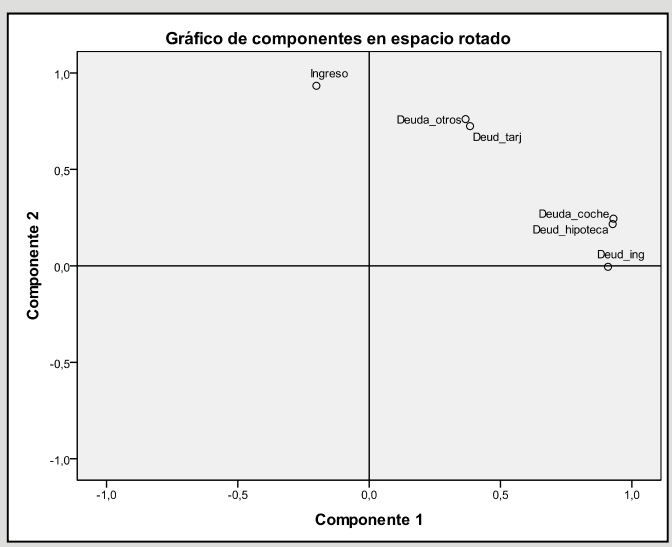
CUADRO XIII.2. *(Continuación)*[SPSS]

Paso 11. La matriz rotada de saturaciones indica la importancia de cada variable en la identificación de los factores. El primer factor está asociado a deudas por hipoteca y coche, y a la proporción entre deudas e ingresos, mientras el segundo se asocia al ingreso, deudas por tarjeta y otras deudas. Parece por lo tanto que el factor que hemos denominado "Capacidad" es el segundo, y "Sostenibilidad" el primero. Las saturaciones nos permiten comprobar la coherencia de nuestro modelo de factores.

Matriz de componentes rotados[a]

	Componente	
	1	2
Ingreso	-,201	,934
Deud_ing	,910	-,005
Deud_tarj	,384	,725
Deud_hipoteca	,930	,245
Deuda_coche	,927	,216
Deuda_otros	,367	,761

Paso 12. Para ello sirve de ayuda el gráfico se saturaciones, que se muestra a continuación. Las variables que se alejan del origen en sentido horizontal son las que identifican el primer factor (deuda_ingresos, deuda_coche y deuda_hipoteca). Las que se alejan en sentido vertical (Ingreso) identifican el segundo factor.



Por tanto, como conclusión, las comunalidades elevadas, el valor moderadamente satisfactorio del índice KMO, y el alto porcentaje de varianza total explicada avalan la adecuación del análisis factorial. Por otra parte la interpretación de las saturaciones y del gráfico correspondiente permiten identificar los factores y estudiar la coherencia del modelo, que puede ser por lo tanto útil para explicar los datos observados.

XIII.3. Análisis de Correspondencias

Es una técnica utilizada para representar las filas y las columnas de una tabla de contingencia de dos variables cualitativas como puntos en un espacio de dos dimensiones, de forma que se superponen las filas y las columnas para obtener una representación conjunta. En general es útil en grandes tablas de contingencia. Por ejemplo, si tenemos veinte especies de árboles y 15 bosques y hemos hecho un recuento por muestreo, tendremos una tabla de contingencia con las especies en filas y los bosques en columnas. Cada casilla de la tabla nos dará la frecuencia relativa bidimensional para cada especie y bosque. El análisis de correspondencias nos permitirá representar las especies y los bosques en un mismo gráfico, de modo que podremos asociar especies con perfil de distribución similar, bosques afines, e identificar las especies características de cada bosque.

1. Perfil de filas y de columnas

Cada fila de la tabla establece un perfil de fila, definido por sus frecuencias.

- Perfil de la fila i-ésima es la proporción de cada categoría de la variable *j* en la fila *i*:

$$r_{ij} = \frac{n_{ij}}{n_i}$$

El perfil de los totales de columna, es decir las frecuencias marginales de todas las columnas, define el *perfil medio* de las filas.

- Perfil de la columna j-ésima es la proporción de cada categoría de la variable *i* en la columna *j*:

$$c_{ij} = \frac{n_{ij}}{n_j}$$

El perfil de los totales de fila, frecuencias marginales de fila, es el *perfil medio* de las columnas**.**

2. Masa de cada elemento o perfil

La *masa* asociada a cada elemento fila (r_i) o columna (c_j) mide la importancia relativa de ese elemento en el conjunto, y es igual a su frecuencia marginal dividida por el total general o tamaño de muestra *n*, es decir, el correspondiente elemento del perfil medio de columnas, mientras que la masa de cada perfil columna será el correspondiente elemento del perfil medio de las filas.

$$r_i = \frac{n_i}{n} = f_{i.} \qquad\qquad c_j = \frac{n_j}{n} = f_{.j}$$

Al representar los perfiles fila/columna deberemos dar más importancia a los que tengan mayor masa.

La mejor medida de distancia entre perfiles (que vienen definidos por las frecuencias) no es la habitual distancia euclídea, sino la distancia Ji-cuadrado, ya que esta distancia tiene una propiedad importante de equivalencia distribucional, según la cual si agregamos dos perfiles iguales, el perfil agregado que sustituye a ambos conserva la distancia de los anteriores con cualquier otro perfil, y no cambia ninguna otra distancia. Como la agregación de filas y columnas es habitual en la construcción de tablas de contingencia, y a menudo se utilizan criterios subjetivos para decidir el número y los límites de las categorías o perfiles, esta propiedad adquiere un interés especial. Por esta razón se realizan algunas transformaciones simples en los datos para obtener esta distancia Ji cuadrado.

Distancia Ji cuadrado entre perfiles fila:

$$d^2_{\chi^2}(r_i, r_{i'}) = \sum_{j=1}^{J} \frac{(r_{ij} - r_{i'j})^2}{c_j}$$

Distancia Ji cuadrado entre perfiles columna:

$$d^2_{\chi^2}(c_j, c_{j'}) = \sum_{i=1}^{I} \frac{(c_{ij} - c_{ij'})^2}{r_i}$$

Para construir el gráfico se utiliza el método de Componentes Principales, con el cual se obtienen generalmente dos dimensiones.

3. Contribuciones absolutas y relativas

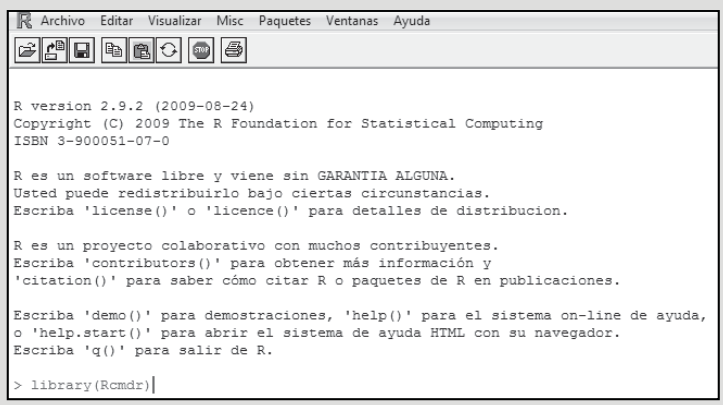
- Contribuciones absolutas: Parte de la variabilidad de cada eje o dimensión (componente principal) que es explicada por cada perfil. Sirven para interpretar los ejes en función de los perfiles, observando cuales son los perfiles que tienen más peso –o contribución absoluta– en la construcción de cada dimensión.
- Contribuciones relativas: Parte de la variabilidad de cada perfil explicado por cada eje o componente. Indican hasta qué punto cada perfil está adecuadamente explicado o representado por los dos ejes o dimensiones. Los perfiles fila o columna que tienen pequeñas contribuciones relativas en ambos ejes no responden al modelo, y están deficientemente representados en el gráfico, lo que debe tenerse en cuenta en la interpretación del mismo.

CUADRO XIII.3. Análisis de correspondencias

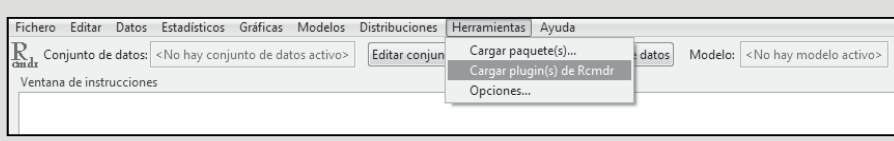
Aplicación con R.

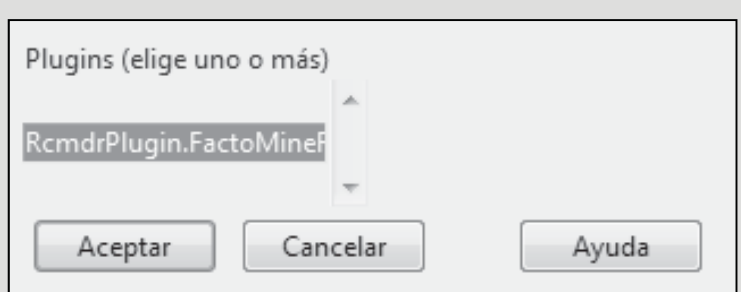
EJEMPLO. En el archivo **Cuadro XIII.3.xls** hay datos de criterios de valoración, en una escala de 0 a 10, de 48 candidatos que optan a un puesto de trabajo. El objetivo es determinar si, con un Análisis de Correspondencias, es posible identificar cuales son las cualidades que permiten diferenciar mejor a los candidatos, para facilitar el proceso de selección.

Paso 1. Iniciamos R y elegimos el directorio de trabajo. Son necesarios los paquetes "Rcmdr", "FactoMineR" y "RcmdrPlugin.FactoMineR". Si los paquetes no están instalados, la instalación se realiza como se explica en el Apéndice II. Se introduce la instrucción «library(Rcmdr)».


Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> library(Rcmdr)

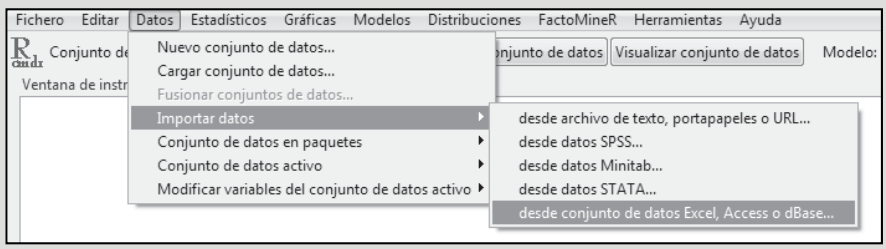
Paso 2. Se selecciona «Herramientas» ⇨ «Cargar plugin(s) de Rcmdr» y en la ventana que aparece se selecciona «RcmdrPlugin.FactoMineR». Será necesario reiniciar Rcmdr para que se cargue definitivamente el paquete "FactoMineR".



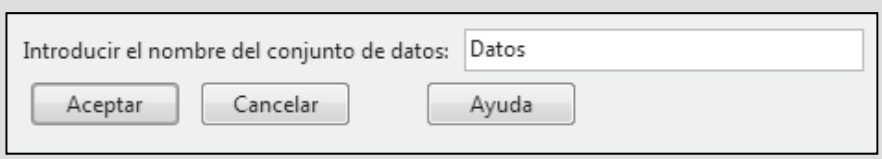


CUADRO XIII.3. *(Continuación)*[R]

Paso 3. En el menú se observa que ahora aparece la opción «FactoMineR», al lado de «Herramientas», que antes no estaba. A continuación se selecciona «Datos» ⇨ «Importar datos» ⇨ «desde conjunto de datos Excel, Access o dBase».


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones FactoMineR Herramientas Ayuda
Nuevo conjunto de datos...
Cargar conjunto de datos...
Fusionar conjuntos de datos...
Importar datos
Conjunto de datos en paquetes
Conjunto de datos activo
Modificar variables del conjunto de datos activo
desde archivo de texto, portapapeles o URL...
desde datos SPSS...
desde datos Minitab...
desde datos STATA...
desde conjunto de datos Excel, Access o dBase...
Visualizar conjunto de datos
Modelo:

Paso 4. Aparece un pequeño cuadro de diálogo para introducir el nombre del conjunto de datos, en el que dejamos el nombre por defecto «Datos» y pulsamos «Aceptar». Ahora aparecen los archivos en el directorio de trabajo, en el que elegimos **Cuadro XIII.3.xls** y pulsamos «Abrir». En R Commander pulsamos en «Visualizar conjunto de datos». Es importante inspeccionar los datos y comprobar que no existan líneas, lo cual a veces ocurre al final del archivo, que tengan celdas con el código <NA>. Si existen estas líneas, hay que eliminarlas.



	Candidato	Carta.de.motivación	Presentación	Nivel.de.estudios	Simpatía	Confianza.en.si.mismo	Lucidez
1	1	6	7	2	5	8	7
2	2	9	10	5	8	10	9
3	3	7	8	3	6	9	8
4	4	5	6	8	5	6	5
5	5	6	8	8	8	4	4
6	6	7	7	7	6	8	7
7	7	9	9	8	8	8	8
8	8	9	9	9	8	9	9
9	9	9	9	7	8	8	8
10	10	4	7	10	2	10	10

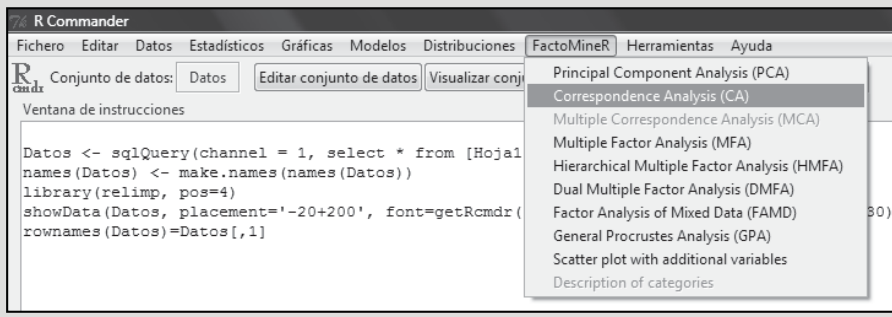
Paso 5. Antes de realizar el análisis de Correspondencias es conveniente asignar etiquetas a las filas, para facilitar la interpretación del gráfico. Para ello utilizamos la orden siguiente, la cual escribiremos en la ventana de instrucciones de R Commander:

```
rownames(Datos)=Datos[,1]
```

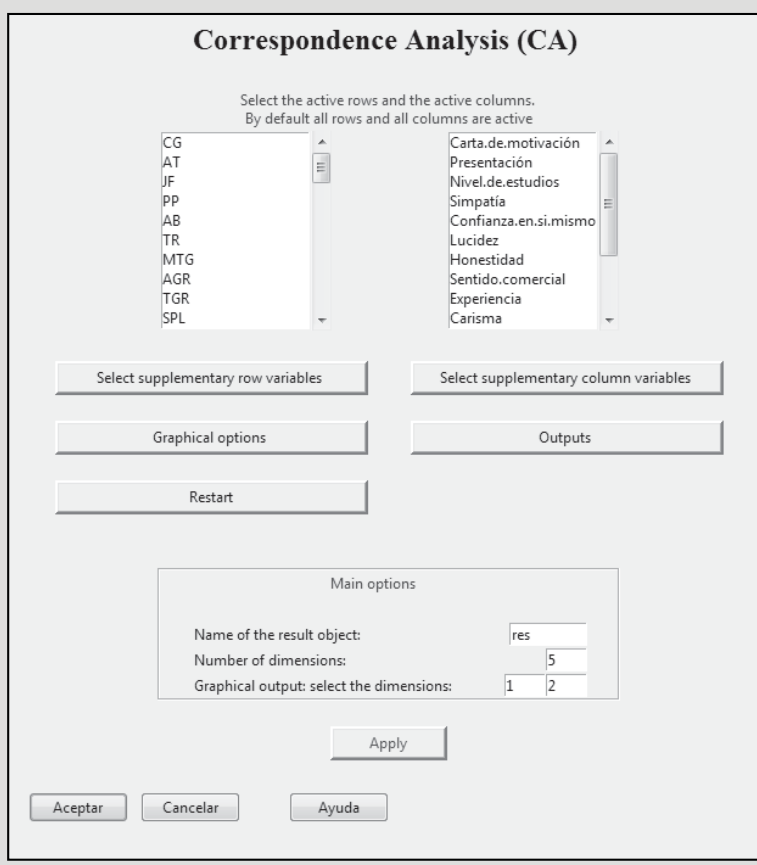
Para ejecutarla, simplemente la seleccionamos y pulsamos «Ejecutar». «Datos» es el nombre que R Commander asigna por defecto a nuestra tabla, y en la columna 1 están las etiquetas que queremos utilizar para cada fila. Las etiquetas de cada fila deben ser letras y no números, y no pueden estar duplicadas.

CUADRO XIII.3. *(Continuación)*[R]

Paso 6. Para realizar el Análisis de Correspondencias utilizaremos la opción de menú «FactoMineR» ⇨ «Correspondence Analysis (CA)».

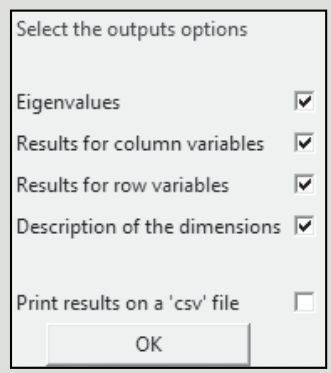

R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones FactoMineR Herramientas Ayuda
Conjunto de datos: Datos
Editar conjunto de datos
Visualizar conj
Ventana de instrucciones
Datos <- sqlQuery(channel = 1, select * from [Hoja1
names(Datos) <- make.names(names(Datos))
library(relimp, pos=4)
showData(Datos, placement='-20+200', font=getRcmdr(
rownames(Datos)=Datos[,1]
Principal Component Analysis (PCA)
Correspondence Analysis (CA)
Multiple Correspondence Analysis (MCA)
Multiple Factor Analysis (MFA)
Hierarchical Multiple Factor Analysis (HMFA)
Dual Multiple Factor Analysis (DMFA)
Factor Analysis of Mixed Data (FAMD)
General Procrustes Analysis (GPA)
Scatter plot with additional variables
Description of categories

Paso 7. Aparece un cuadro de diálogo con los códigos de los casos en la ventana de la izquierda y las variables a la derecha. Por defecto están seleccionados todos los casos y todas las variables. En el caso de no querer trabajar con todos los datos, se seleccionan las filas y columnas que queremos utilizar.

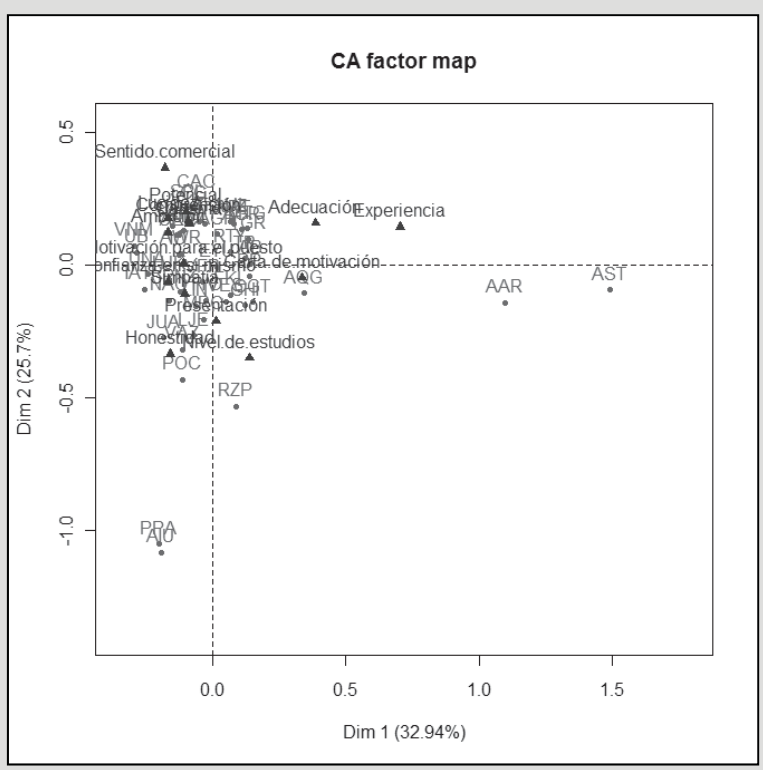


CUADRO XIII.3. *(Continuación)*[R]

Paso 8. En la ventana anterior se pulsa en «Outputs» y se marcan todas las opciones. También se pueden guardar los resultados en un archivo "csv". Posteriormente se pulsa «Apply».



Paso 9. El gráfico muestra que en el extremo derecho del primer eje está la experiencia junto con la adecuación al puesto y carta de motivación. En el extremo izquierdo del primer eje están asociadas la motivación y confianza. El segundo eje viene asociado al sentido comercial en el extremo superior, mientras que en el inferior están a la honestidad y nivel de estudios.



CUADRO XIII.3. *(Continuación)*[R]

Paso 10. Los autovalores «res$eig» muestran el porcentaje de varianza de los ejes: 32,9% primer eje, 25,69% el segundo, etc.

```
> res$eig
        eigenvalue percentage of variance cumulative percentage of variance
dim 1  5.474888e-02           3.293708e+01                          32.93708
dim 2  4.271094e-02           2.569502e+01                          58.63210
dim 3  1.879606e-02           1.130776e+01                          69.93987
dim 4  1.315684e-02           7.915192e+00                          77.85506
dim 5  8.201061e-03           4.933781e+00                          82.78884
dim 6  6.547529e-03           3.939012e+00                          86.72785
dim 7  6.121793e-03           3.682888e+00                          90.41074
dim 8  4.844037e-03           2.914186e+00                          93.32493
dim 9  3.893567e-03           2.342381e+00                          95.66731
dim 10 2.278127e-03           1.370528e+00                          97.03783
dim 11 2.016690e-03           1.213247e+00                          98.25108
dim 12 1.380054e-03           8.302442e-01                          99.08132
dim 13 9.083971e-04           5.464943e-01                          99.62782
dim 14 6.186492e-04           3.721811e-01                         100.00000
dim 15 2.973552e-33           1.788898e-30                         100.00000
```

Paso 11. Los coordenadas de las variables en los ejes «$coord» permiten poder ubicar cada variable en la gráfica mostrada anteriormente y poder disponer de las coordenadas para realizar cualquier otro tipo de representación.

```
> res$col
$coord
                              Dim 1       Dim 2       Dim 3       Dim 4
Carta.de.motivación         0.33592135 -0.048766522  0.203423410  0.19066673
Presentación                0.01140477 -0.211711595 -0.061786673  0.05421962
Nivel.de.estudios           0.13785646 -0.352685521 -0.214438919  0.03196260
Simpatía                   -0.10513156 -0.110333413  0.227640854 -0.18086951
Confianza.en.si.mismo      -0.17086368 -0.064726705 -0.079078785  0.10769042
Lucidez                    -0.16898816  0.174297329 -0.038545318 -0.06174628
Honestidad                 -0.16206340 -0.336935372  0.005620678 -0.11737064
Sentido.comercial          -0.18049741  0.363764566 -0.128264153  0.04787293
Experiencia                 0.70255602  0.142384506 -0.055688670 -0.11105826
Carisma                    -0.09108127  0.152641306 -0.021855650  0.10725045
Ambición                   -0.16981181  0.120217510 -0.054501161  0.16368188
Comprensión                -0.09186391  0.165493699 -0.041498283 -0.10328221
Potencial                  -0.10052240  0.199879290 -0.027449090 -0.15099507
Motivación.para.el.puesto -0.11206197  0.003851038  0.333133717  0.08526804
Adecuación                  0.38635737  0.157190735 -0.009016265 -0.05664184
```

Paso 12. Las contribuciones absolutas de las variables «$contrib» ayudan a entender mejor los ejes, ya que indican el peso que cada variable tiene en la formación de cada uno de los ejes. En el primer eje los candidatos se diferencian preferentemente en función de la experiencia (41,69%), adecuación (17,76%) y carta de motivación (13,52%). La segunda dimensión, o eje vertical en el gráfico, viene determinada fundamentalmente por la honestidad (23,37%), el nivel de estudios (22,56%) y sentido comercial (16,44%). Por lo tanto, estas son las variables que mejor diferencian a los distintos candidatos y, por ello, podrían ser utilizadas para elegir entre ellos.

CUADRO XIII.3. *(Continuación)*[R]

```
$contrib
                              Dim 1        Dim 2        Dim 3       Dim 4
Carta.de.motivación       13.52467602  0.365368723 14.44647662 18.1311423
Presentación               0.01840394  8.129494079  1.57338909  1.7309106
Nivel.de.estudios          2.68900926 22.560540161 18.95194232  0.6015142
Simpatía                   1.35689870  1.915717358 18.53062289 16.7122747
Confianza.en.si.mismo      4.04578588  0.744228932  2.52424674  6.6877687
Lucidez                    3.60092728  4.910428674  0.54569941  2.0005392
Honestidad                 4.21906658 23.376303223  0.01478197  9.2085046
Sentido.comercial          3.15905381 16.447201818  4.64658732  0.9247381
Experiencia               41.69818148  2.195415974  0.76312787  4.3359135
Carisma                    0.88035424  3.169418890  0.14765060  5.0795039
Ambición                   3.44410884  2.212649710  1.03338224 13.3157603
Comprensión                1.05358661  4.383080968  0.62625196  5.5418551
Potencial                  1.14801484  5.818265810  0.24933666 10.7788159
Motivación.para.el.puesto  1.39536203  0.002112332 35.91832132  3.3617600
Adecuación                17.76657047  3.769773348  0.02818298  1.5889989
```

Paso 13. Las contribuciones relativas de las columnas «$cos2» indican el peso de cada eje en la explicación de cada una de las variables o perfiles de columna. Así, la experiencia está explicada por el primer eje en un 84,1% frente al 3,4% del segundo, la adecuación está explicada por el primer eje en un 65,1% y 10,7% en el segundo, etc.

```
$cos2
                               Dim 1         Dim 2         Dim 3        Dim 4
Carta.de.motivación       0.49293249 1.038857e-02 0.1807648911 0.158804227
Presentación              0.00143008 4.928071e-01 0.0419737028 0.032322172
Nivel.de.estudios         0.08755395 5.730557e-01 0.2118501458 0.004706582
Simpatía                  0.08238209 9.073621e-02 0.3862486282 0.243835673
Confianza.en.si.mismo     0.40738189 5.846140e-02 0.0872614059 0.161829016
Lucidez                   0.25261310 2.687353e-01 0.0131427598 0.033726005
Honestidad                0.14838968 6.413971e-01 0.0001784889 0.077831024
Sentido.comercial         0.14383341 5.841961e-01 0.0726320733 0.010118074
Experiencia               0.84108034 3.454626e-02 0.0052845652 0.021017318
Carisma                   0.06989568 1.963073e-01 0.0040245716 0.096914884
Ambición                  0.30110996 1.509123e-01 0.0310170735 0.279763170
Comprensión               0.10082424 3.272187e-01 0.0205747895 0.127446002
Potencial                 0.08728958 3.451217e-01 0.0065086755 0.196952856
Motivación.para.el.puesto 0.08184850 9.666068e-05 0.7233215970 0.047387833
Adecuación                0.65131440 1.078118e-01 0.0003547036 0.013998677
```

Paso 14. La inercia «$inertia» indica el peso relativo o la distinta importancia de cada variable. Los mayores valores son los de la experiencia (0,0271), nivel de estudios (0,0168), honestidad (0,0155), carta de motivación (0,015), adecuación (0,0149), sentido comercial (0,012), etc.

```
$inertia
      Carta.de.motivación              Presentación         Nivel.de.estudios
                0.015021548               0.007045725                0.016814803
                   Simpatía     Confianza.en.si.mismo                   Lucidez
                0.009017578               0.005437214                0.007804296
                 Honestidad         Sentido.comercial                Experiencia
                0.015566391               0.012024651                0.027142815
                    Carisma                 Ambición                Comprensión
                0.006895764               0.006262201                0.005721113
                  Potencial Motivación.para.el.puesto                Adecuación
                0.007200462               0.009333648                0.014934414
```

CUADRO XIII.3. *(Continuación)*[R]

Paso 15. Al igual que para las variables, también se presentan los datos para los casos, en este caso los candidatos, es decir, las coordenadas «$coord», las contribuciones absolutas «$contrib», la inercia «$inertia», etc. De especial interés para seleccionar al candidato son las contribuciones absolutas y las coordenadas de los casos.

Si para seleccionar al candidato nos basamos en las variables que más influían en el primer eje en su extremo derecho (experiencia y adecuación), los candidatos que tienen las coordenadas más positivas en este primer eje son AST (1,495) y AAR (1,098).

```
ART   0.07641991  0.155705636  0.05273385 -0.037905933  0.053144274
AVZ   0.06527882  0.161622207  0.04648016 -0.017559978  0.053650786
AAR   1.09862897 -0.142754653 -0.14780338  0.109127735 -0.091971615
AST   1.49538258 -0.094406921 -0.10708812  0.140700774 -0.132361863
AQG   0.34219808 -0.108162398 -0.01443342 -0.285189941  0.080285802
```

que además son los candidatos que tienen más peso en el primer eje: AST (40,01%) y AAR (27,62%).

```
ART   0.345112394 1.836506e+00  0.478668594 3.533341e-01 1.114207e+00
AVZ   0.255367856 2.006596e+00  0.377107572 7.689415e-02 1.151540e+00
AAR  27.626377931 5.979138e-01  1.456461430 1.134268e+00 1.292512e+00
AST  40.015869065 2.044426e-01  0.597748685 1.474159e+00 2.092949e+00
AQG   3.118848512 3.994182e-01  0.016161653 9.014284e+00 1.146098e+00
```

Si por el contrario, nos basamos para la selección en las variables que mejor discriminaban el segundo eje, como por ejemplo sentido comercial que estaba en el extremo positivo superior del eje, entonces los candidatos seleccionados deberían ser CAC (0,252) o SPL (0,216).

```
TGR   0.13403225  0.095758418  0.05465795  0.004418798  0.034071276
SPL  -0.09451866  0.216408059 -0.27976530  0.067503018  0.027074537
CAC  -0.06011790  0.252123983 -0.36508836  0.150074947  0.002096545
SQR  -0.15167828  0.144950865 -0.23577674  0.001700287 -0.051895422
FKL   0.06447852 -0.114655454  0.09405076 -0.103383832  0.063100513
GHI   0.12212444 -0.153248509  0.09110775 -0.066839005 -0.076170461
CGT   0.14875694 -0.141148784 -0.03006603 -0.192312896  0.084510762
HJL   0.07208037  0.165423159 -0.04309447 -0.031217734  0.089574056
```

que lógicamente son los candidatos que tienen más peso en el segundo eje: CAC (3,39%) y SPL (2,87%).

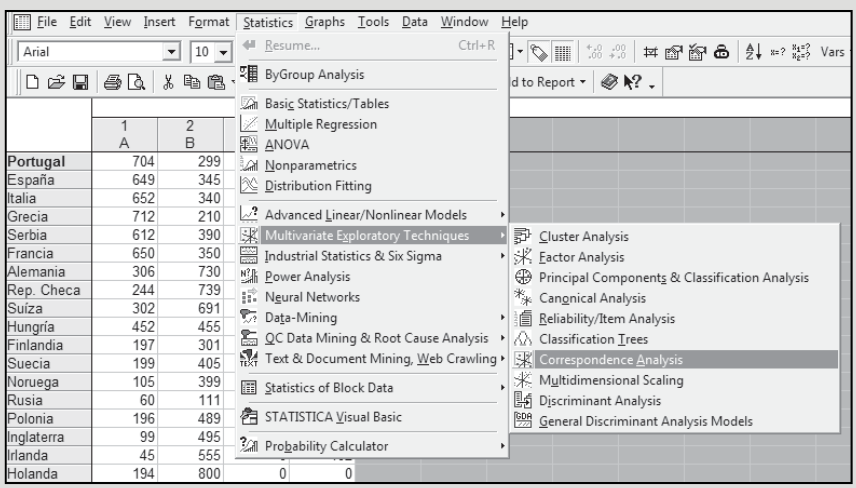
```
TGR   0.912089834 5.967729e-01  0.441808520 4.125256e-03 3.934604e-01
SPL   0.427555084 2.873029e+00 10.910713076 9.074591e-01 2.341986e-01
CAC   0.150406553 3.390968e+00 16.157105655 3.900311e+00 1.221160e-03
SQR   1.034021856 1.210489e+00  7.277685392 5.406937e-04 8.080649e-01
FKL   0.160906009 6.521803e-01  0.997183606 1.721357e+00 1.028758e+00
GHI   0.558607595 1.127536e+00  0.905567299 6.962822e-01 1.450708e+00
CGT   0.782767095 9.033759e-01  0.093140531 5.444006e+00 1.686582e+00
HJL   0.248651443 1.678749e+00  0.258886004 1.940812e-01 2.563461e+00
```

CUADRO XIII.3. *(Continuación)*[STATISTICA]

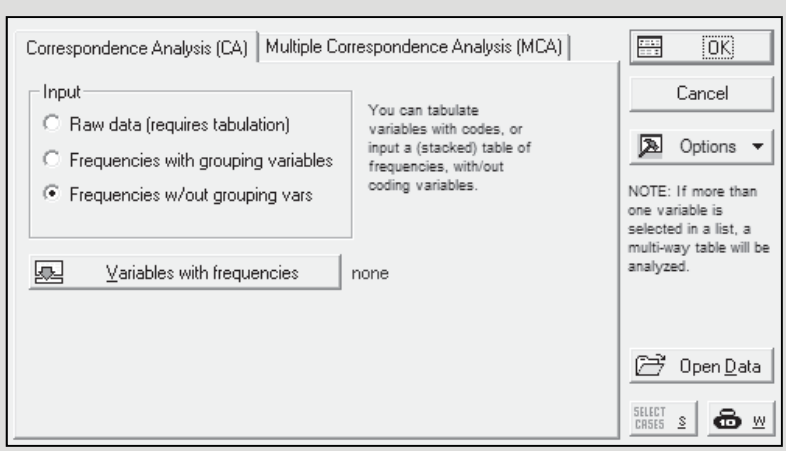
Aplicación con STATISTICA.

EJEMPLO. Se ha realizado una encuesta para estudiar la frecuencia de cuatro marcas de automóvil (identificadas como A, B, C y D) en distintos países europeos. El tamaño de muestra es aproximadamente 1000 automóviles elegidos al azar en cada país. Los datos se encuentran en el archivo **Cuadro XIII.3.sta**. El objetivo es determinar el grado de asociación que existe entre los países en su preferencia por las marcas de automóviles.

Paso 1. Seleccionamos en el menú «Statistics» ⇨ «Multivariate Exploratory Techniques» ⇨ «Correspondence Analysis».



Paso 2. Aparece un cuadro de diálogo en el que debemos marcar la opción «Frequencies w/out grouping vars», ya que nuestros datos son ya una tabla de frecuencias.


Correspondence Analysis (CA)
Multiple Correspondence Analysis (MCA)
Input
Raw data (requires tabulation)
Frequencies with grouping variables
Frequencies w/out grouping vars
You can tabulate variables with codes, or input a (stacked) table of frequencies, with/out coding variables.
Variables with frequencies
none
OK
Cancel
Options
NOTE: If more than one variable is selected in a list, a multi-way table will be analyzed.
Open Data

CUADRO XIII.3. *(Continuación)*[STATISTICA]

La segunda opción «Frequencies with grouping variables» se utilizaría si cada caso fuese marca de automóvil y país, indicando con otra variable la frecuencia correspondiente.

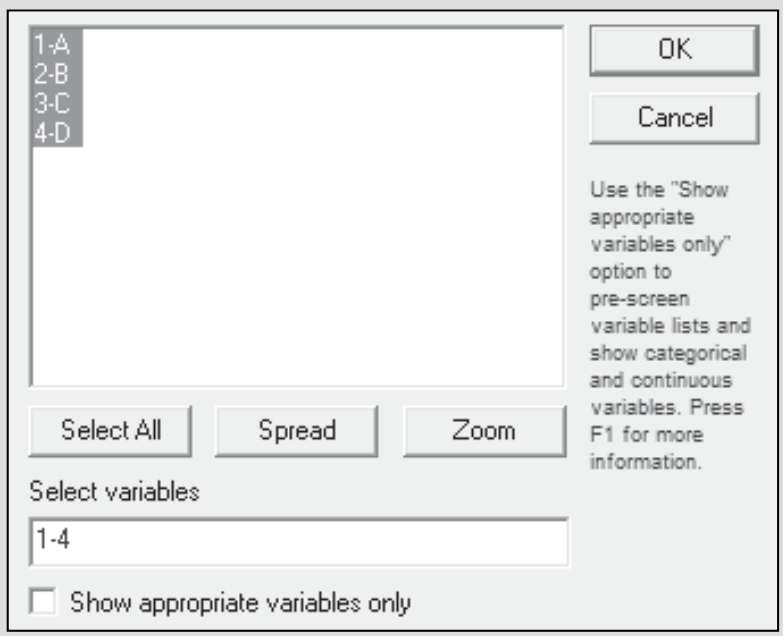
	1 Var1	2 Var2	3 Var3
1	Portugal	A	704
2	Portugal	B	299
3	Portugal	C	0
4	Portugal	D	0
5	España	A	649
6	España	B	345
7	España	C	0
8	España	D	0

Si tuviésemos datos brutos, es decir cada caso o fila un automóvil, con su marca y país como variables, utilizaríamos la primera opción o datos brutos «Raw data».

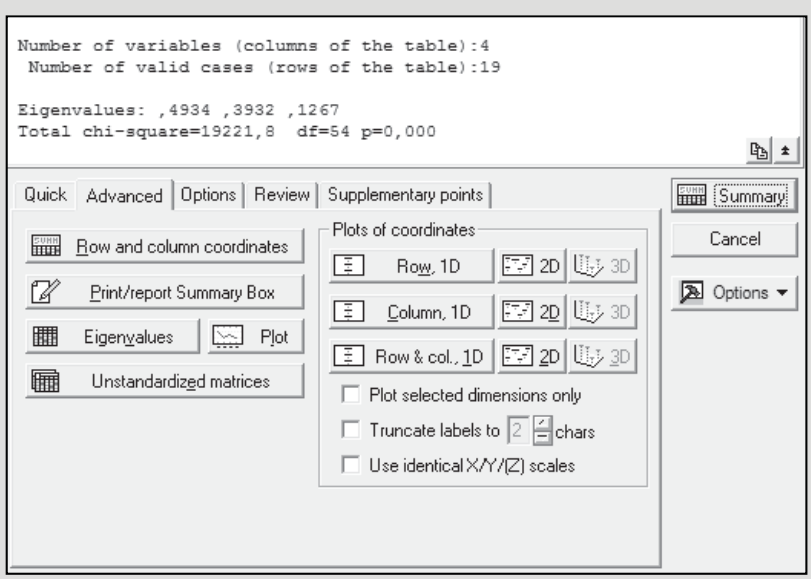
	1 Var1	2 Var2
1	Italia	A
2	Francia	A
3	Finlandia	A
4	Portugal	B
5	Portugal	C
6	Portugal	C
7	Portugal	C
8	Rusia	B
9	Alemania	A
10	España	B

CUADRO XIII.3. *(Continuación)*[STATISTICA]

Paso 3. Se pulsa en «Variables with frequencies» para que abra un nuevo cuadro de diálogo en el que seleccionaremos las variables o columnas a utilizar (en este caso todas). Pulsamos sobre «OK» sucesivamente en este cuadro y el anterior.



Paso 4. En la pestaña «Advanced» pulsamos «Eigenvalues» (Autovalores).



CUADRO XIII.3. *(Continuación)*[STATISTICA]

Paso 5. Las dos dimensiones son componentes principales obtenidas sobre la matriz de frecuencias transformada. Los autovalores representan la varianza (o inercia) explicada por esas componentes: 48,69% la primera, 38,80% la segunda y 87,49% entre las dos componentes.

	Eigenvalues and Inertia for all Dimensions (Cuadro XIII.3.sta) Input Table (Rows x Columns): 19 x 4 Total Inertia=1,0133 Chi²=19222, df=54 p=0,0000				
Number of Dims.	Singular Values	Eigen-Values	Perc. of Inertia	Cumulatv Percent	Chi Squares
1	0,702439	0,493421	48,69323	48,6932	9359,703
2	0,627055	0,393198	38,80271	87,4959	7458,570
3	0,355959	0,126707	12,50405	100,0000	2403,501

Paso 6. En la ventana del paso 4 pulsamos la opción «Row and column coordinates». Aparecen dos tablas, una para las filas, y otra para las columnas. Podemos elegir cualquiera de las dos en la parte izquierda de la ventana: «Row Coordinates» o «Column Coordinates». Ambas tablas se interpretan del mismo modo.

	Row Coordinates and Contributions to Inertia (Cuadro XIII.3.sta) Input Table (Rows x Columns): 19 x 4 Standardization: Row and column profiles									
Row Name	Row Number	Coordin. Dim.1	Coordin. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine² Dim.1	Inertia Dim.2	Cosine² Dim.2
Portugal	1	-0,334426	0,57552	0,052876	0,721988	0,032021	0,011985	0,182250	0,044541	0,539739
España	2	-0,330533	0,53141	0,052401	0,800343	0,025306	0,011603	0,223257	0,037635	0,577085
Italia	3	-0,330878	0,53532	0,052296	0,793053	0,025773	0,011603	0,219224	0,038114	0,573829
Grecia	4	-0,340017	0,63886	0,048606	0,627714	0,040022	0,011389	0,138559	0,050453	0,489155
Serbia	5	-0,327183	0,49346	0,052823	0,872700	0,020940	0,011460	0,266495	0,032713	0,606205
Francia	6	-0,330301	0,52878	0,052718	0,805271	0,025113	0,011656	0,226013	0,037489	0,579258
Alemania	7	-0,302113	0,20943	0,054615	0,432713	0,016832	0,010103	0,292266	0,006092	0,140447
Rep. Checa	8	-0,298365	0,16697	0,051821	0,303773	0,019680	0,009349	0,231327	0,003674	0,072446
Suíza	9	-0,302809	0,21732	0,052349	0,461115	0,015564	0,009728	0,304354	0,006288	0,156761
Hungría	10	-0,056270	0,28744	0,053087	0,986889	0,004554	0,000341	0,036425	0,011155	0,950463
Finlandia	11	1,002242	-0,18064	0,052401	0,993530	0,053981	0,106677	0,962271	0,004349	0,031259
Suecia	12	0,729778	-0,11582	0,052559	0,992903	0,028522	0,056730	0,968510	0,001793	0,024394
Noruega	13	1,004500	-0,26354	0,052823	0,988459	0,056875	0,108020	0,924802	0,009331	0,063657
Rusia	14	1,876506	-0,50720	0,053403	0,981786	0,202825	0,381107	0,914944	0,034939	0,066842
Polonia	15	0,655772	-0,11438	0,056882	0,957552	0,025977	0,049575	0,929281	0,001893	0,028271
Inglaterra	16	-0,922625	-1,56414	0,052243	0,978099	0,173828	0,090128	0,252471	0,325066	0,725628
Irlanda	17	-0,919194	-1,61602	0,052823	0,992598	0,181523	0,090452	0,242637	0,350839	0,749960
Holanda	18	-0,294148	0,11920	0,052401	0,204257	0,025503	0,009189	0,175446	0,001894	0,028811
Bélgica	19	-0,293873	0,11608	0,050872	0,199190	0,025163	0,008904	0,172305	0,001743	0,026885

	Column Coordinates and Contributions to Inertia (Cuadro XIII.3.sta) Input Table (Rows x Columns): 19 x 4 Standardization: Row and column profiles									
Column Name	Column Number	Coordin. Dim.1	Coordin. Dim.2	Mass	Quality	Relative Inertia	Inertia Dim.1	Cosine² Dim.1	Inertia Dim.2	Cosine² Dim.2
A	1	-0,25156	0,52921	0,345986	0,720776	0,162647	0,044373	0,132844	0,246440	0,587932
B	2	-0,19572	-0,03546	0,468343	0,228493	0,080031	0,036361	0,221230	0,001498	0,007263
C	3	1,62956	-0,41567	0,143550	0,993543	0,403258	0,772546	0,932846	0,063079	0,060697
D	4	-1,31100	-2,53605	0,042121	0,956853	0,354064	0,146721	0,201780	0,688983	0,755073

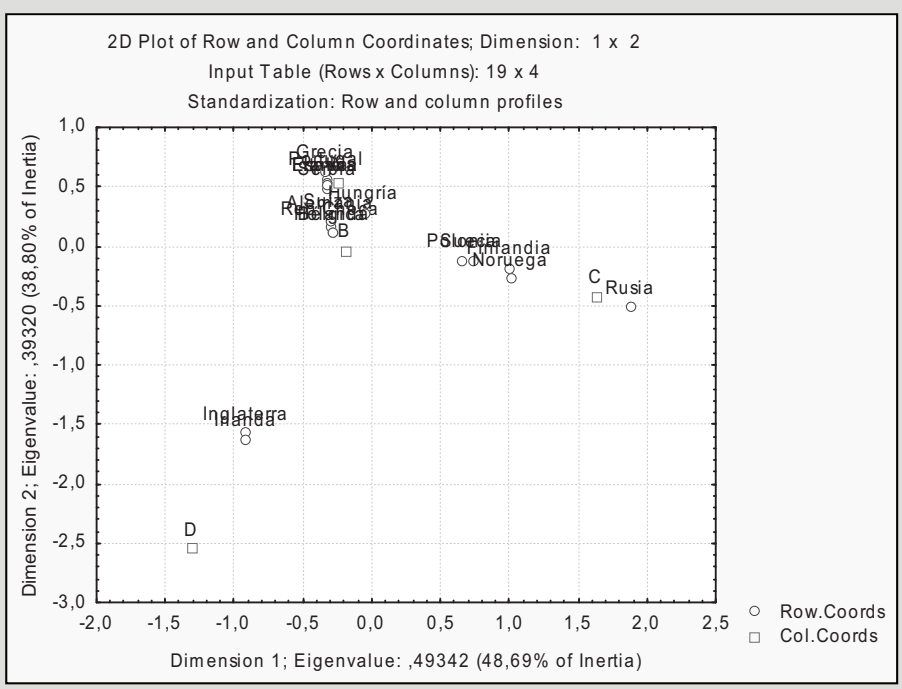
Observemos por ejemplo la segunda, la de las columnas o marcas de automóvil. En primer lugar aparecen las coordenadas, que permitirán la representación gráfica. En segundo lugar la masa, que muestra la diferente importancia o peso de cada marca dentro del conjunto de datos: la marca B supone el 46,83% de los automóviles de la muestra.

CUADRO XIII.3. *(Continuación)*[STATISTICA]

Las cuatro últimas columnas muestran para cada eje (componente, dimensión) la contribución absoluta «Inertia» o porcentaje de cada eje explicado por el perfil correspondiente, y la contribución relativa «cosine», o porcentaje de variabilidad de cada perfil explicado por la componente. La columna «Quality» indica la suma de contribuciones relativas, y es la parte de variabilidad de cada perfil explicado por ambos ejes; cuanto más próximo a 1 mejor será la representación de ese perfil. Las marcas C (99,3%) y D (95,6%) están bien explicadas, pero la marca B solamente tiene un 22,8% y, por tanto, la calidad de representación es muy baja para esta marca B.

Por último, la columna «Relative Inertia» muestra la media de contribuciones absolutas, e indica la distinta importancia de cada perfil en la definición de ambos ejes. Todos estos valores deben tenerse en cuenta en la interpretación del gráfico.

Paso 7. El gráfico de los perfiles fila y columna es el resultado principal del Análisis de Correspondencias. Para obtenerlo, en la ventana del paso 4 pulsamos en "Row & Column 2D".



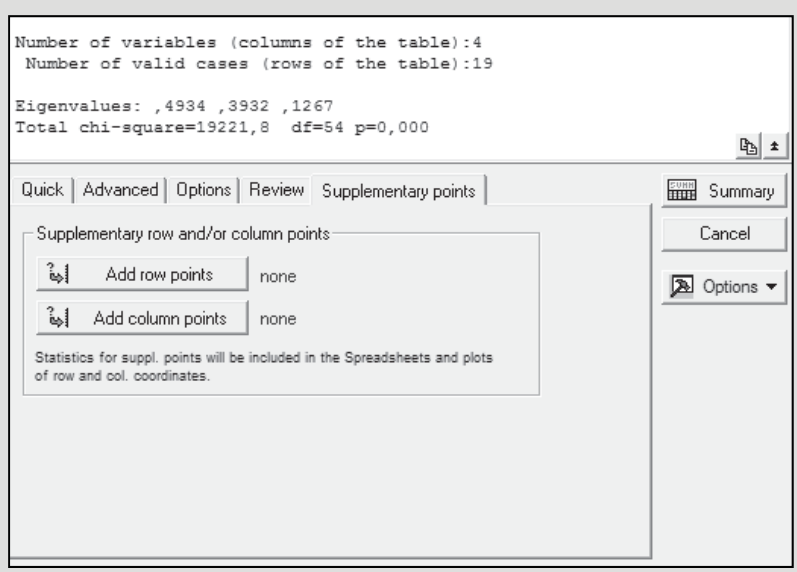
Observamos como los países forman grupos: Inglaterra e Irlanda; Polonia, Finlandia, Suecia y Noruega, los países mediterráneos y centroeuropeos. La proximidad entre países en el gráfico indica un perfil similar de marcas de automóvil.

CUADRO XIII.3. *(Continuación)*[STATISTICA]

Como hemos representado también (puntos rojos) los perfiles columna, o marcas de automóvil, podemos ver como las marcas A y B tienen perfiles de distribución similares, distintos de C o D. Además la representación conjunta nos permite relacionar los perfiles fila y columna: Inglaterra e Irlanda se parecen porque comparten la marca D. Rusia está separada por el predominio de la marca C. Los países mediterráneos y centroeuropeos comparten las marcas A y B, y los países bálticos las marcas B y C.

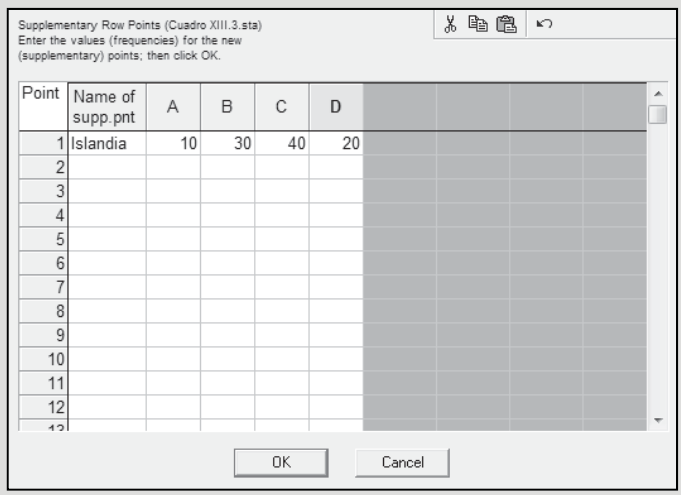
Debemos recordar que no todas las marcas -ni todos los países- tienen igual peso, y que las dimensiones también tienen diferente importancia. El eje horizontal tiene una inercia de 48,69% y el vertical 38,80%. Por tanto, en este caso las diferencias entre el primer eje y el segundo no son muy importantes, pero es posible que en otros casos las diferencias sean mayores, de modo que solo la variación observada en sentido horizontal sea importante.

Paso 8. Una opción a mayores que tenemos ahora es añadir a nuestro análisis nuevos puntos (ya sean de variables en filas o en columnas) que no estaban en el análisis original. Esto podría suceder, por ejemplo, en una situación en la que nos sorprendemos por algunos de los resultados obtenidos, pero que son poco claros y queremos incidir con nuevos puntos que nos den información adicional para ver si esa pauta se refuerza. Pensemos, por ejemplo, que nos sorprendiese la situación tan alejada de Inglaterra e Irlanda, y queremos comprobar si esto pudiera darse en otras grandes islas del continente. Entonces, queremos introducir los datos correspondientes a Islandia. Tendríamos que ir en la ventana del paso 4 a la pestaña «Suplementary points», y aquí añadir los datos correspondientes a una nueva fila «Add row points» o si quisiésemos hacerlo para una columna sería en "Add column points».



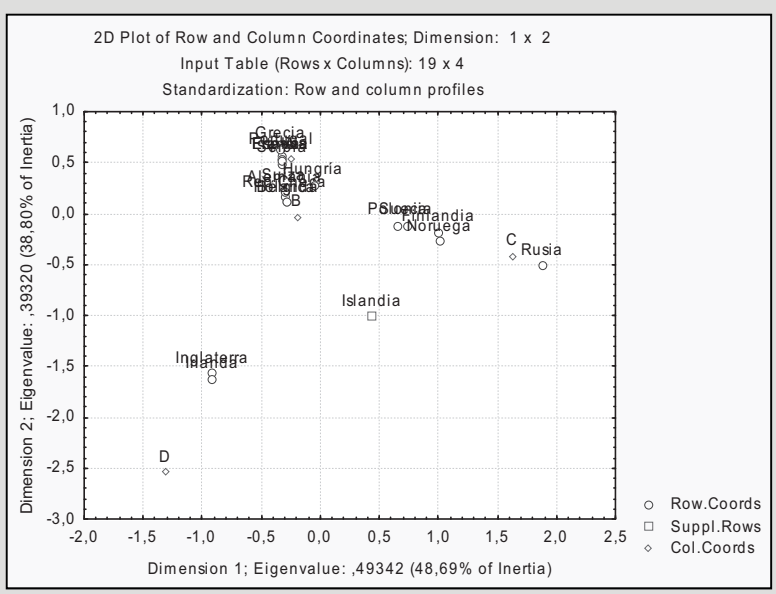
CUADRO XIII.3. *(Continuación)* STATISTICA

Paso 9. Se introducen los siguientes datos para Islandia (hay que introducir frecuencias relativas, en porcentaje o en tanto por uno).


Supplementary Row Points (Cuadro XIII.3.sta)
Enter the values (frequencies) for the new (supplementary) points; then click OK.
OK
Cancel

Point	Name of supp.pnt	A	B	C	D
1	Islandia	10	30	40	20
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					

Paso 10. Repitiendo todos los pasos se obtienen de nuevo los resultados. Para resumir, vamos a mostrar sólo el gráfico en dos dimensiones. Como vemos, para Islandia incluye un nuevo símbolo y lo representa sin asociarse a ninguno de los grupos anteriores. Aunque es de los que más se aproxima al grupo de las islas, parece situarse al margen del resto de tendencias continentales. Al mismo tiempo, es el que tiene preferencias más compensadas entre las distintas marcas.

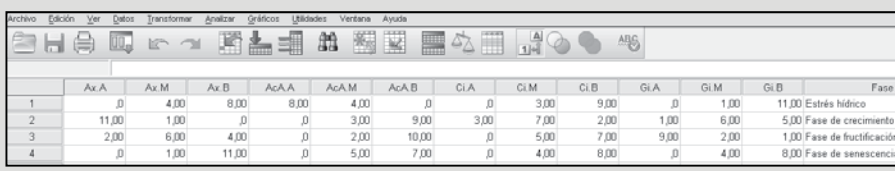


CUADRO XIII.3. *(Continuación)*[SPSS]

Aplicación con SPSS.

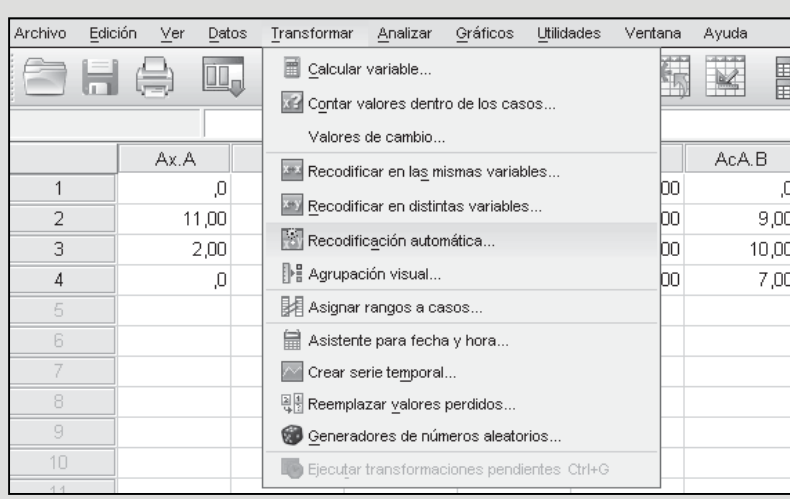
EJEMPLO. En el archivo **Cuadro XIII.3.sav** hay datos de los niveles (alto, medio y bajo) de cuatro clases de hormonas (auxinas, ácido abscísico, citoquininas y giberelina) medidas en 12 individuos de una especie vegetal que estaba en diferentes fases fisiológicas. Auxina con altas concentraciones es «Au.A», Giberelina a bajas concentraciones es «Gi.B», etc. El objetivo es determinar el grado de asociación que existe entre las hormonas y la fase en la que se encuentra el individuo.

Paso 1. Para realizar el Análisis de Correspondencias con SPSS es mejor construir una tabla de frecuencias a partir de los datos brutos como se muestra en la siguiente ventana. En primer lugar hay que poner las variables que indican las frecuencias (en este caso las diferentes hormonas con sus niveles) y luego la variable con la etiqueta o identificador de los casos (en este caso las fases).



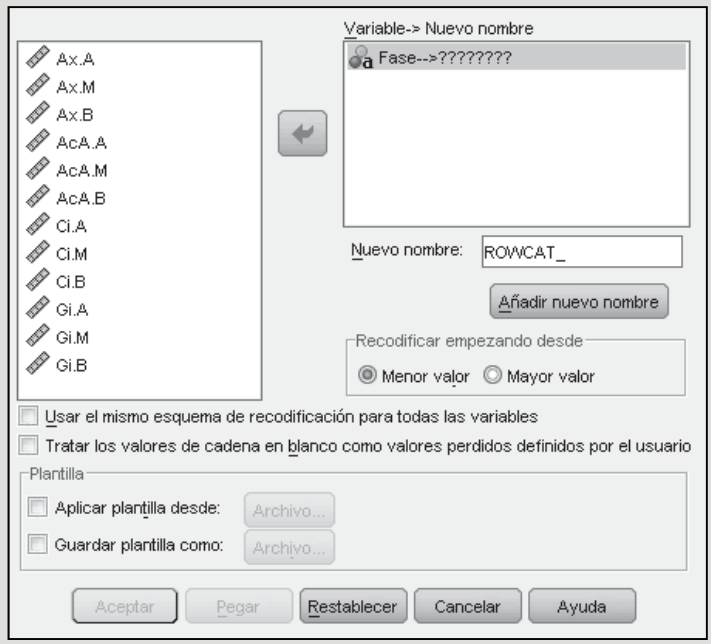
	Ax.A	Ax.M	Ax.B	AcA.A	AcA.M	AcA.B	Ci.A	Ci.M	Ci.B	Gi.A	Gi.M	Gi.B	Fase
1	,0	4,00	8,00	8,00	4,00	,0	,0	3,00	9,00	,0	1,00	11,00	Estrés hídrico
2	11,00	1,00	,0	,0	3,00	9,00	3,00	7,00	2,00	1,00	6,00	5,00	Fase de crecimiento
3	2,00	6,00	4,00	,0	2,00	10,00	,0	5,00	7,00	9,00	2,00	1,00	Fase de fructificación
4	,0	1,00	11,00	,0	5,00	7,00	,0	4,00	8,00	,0	4,00	8,00	Fase de senescencia

Paso 2. En el caso de que la variable con la etiqueta de los casos no sea numérica, debe ser transformada utilizando la opción de «Transformar» ⇨ «Recodificación automática», que convierte los valores de texto, es decir, asocia las fases fisiológicas a un valor numérico (con números enteros consecutivos). Además es necesario que esta variable se denomine «ROWCAT_», palabra reservada en SPSS para designar nombres de fila; de ese modo esta variable no formará parte de los datos con los que trabajará SPPS para realizar el análisis. Se selecciona «Transformar» ⇨ «Recodificación automática…».

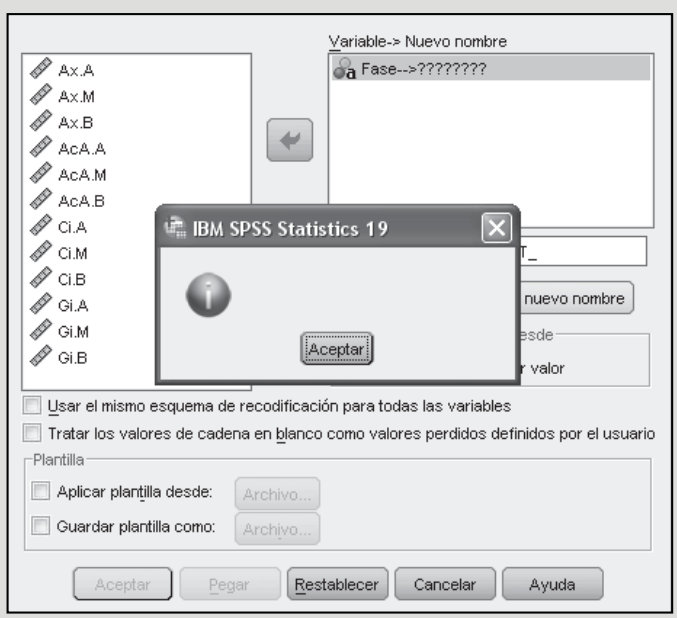


CUADRO XIII.3. *(Continuación)*[SPSS]

Paso 3. En «Variable→Nuevo nombre» introducimos la variable «Fase» y en «Nuevo nombre:» introducimos «ROWCAT_».

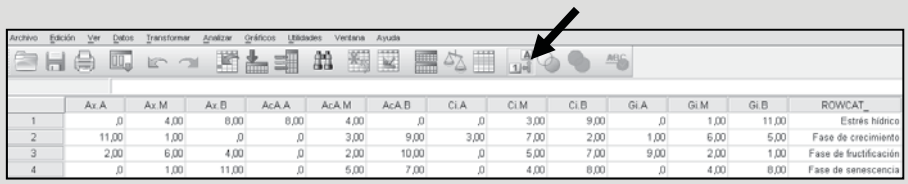


Paso 4. Al pulsar en «Añadir nuevo nombre» aparece esta nueva ventana y pulsamos «Aceptar».



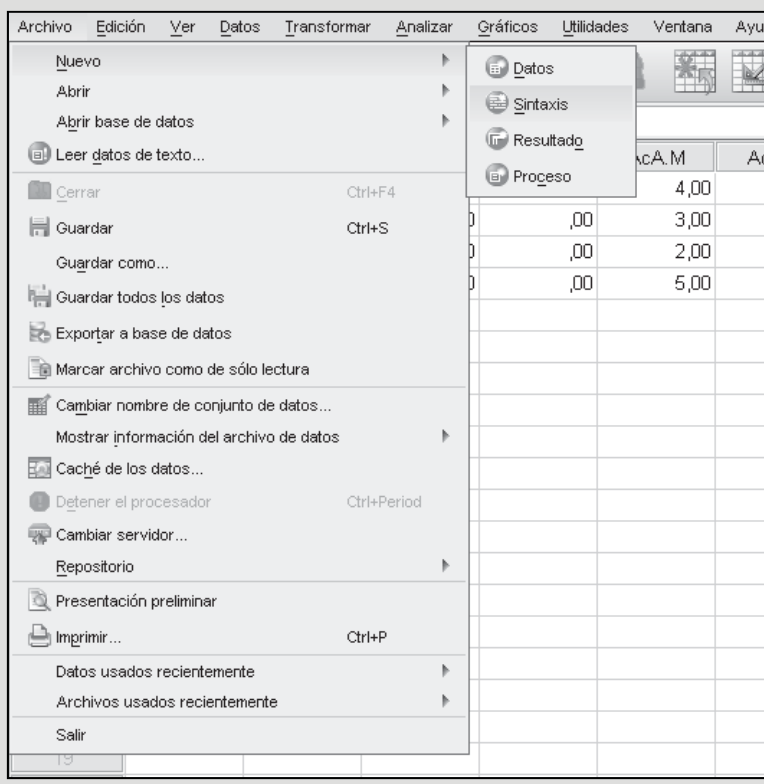
CUADRO XIII.3. *(Continuación)*[SPSS]

Paso 5. A continuación es mejor eliminar la variable «Fase» y quedarnos sólo con «ROWCAT_». Pulsando en el icono que se muestra con la flecha es posible mostrar las etiquetas asignadas en «ROWCAT_» a cada uno de los casos.



	Ax.A	Ax.M	Ax.B	AcA.A	AcA.M	AcA.B	Ci.A	Ci.M	Ci.B	Gi.A	Gi.M	Gi.B	ROWCAT_
1	,0	4,00	8,00	8,00	4,00	,0	,0	3,00	9,00	,0	1,00	11,00	Estrés hídrico
2	11,00	1,00	,0	,0	3,00	9,00	3,00	7,00	2,00	1,00	6,00	5,00	Fase de crecimiento
3	2,00	6,00	4,00	,0	2,00	10,00	,0	5,00	7,00	9,00	2,00	1,00	Fase de fructificación
4	,0	1,00	11,00	,0	5,00	7,00	,0	4,00	8,00	,0	4,00	8,00	Fase de senescencia

Paso 6. Para aplicar el Análisis de correspondencias creamos un archivo de sintaxis con «Archivo» ⇨ «Nuevo» ⇨ «Sintaxis».

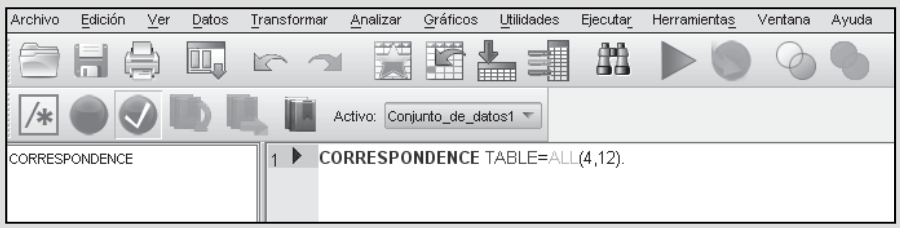


Paso 7. En la ventana de sintaxis escribimos la instrucción:

CORRESPONDENCE TABLE=ALL(4,12).

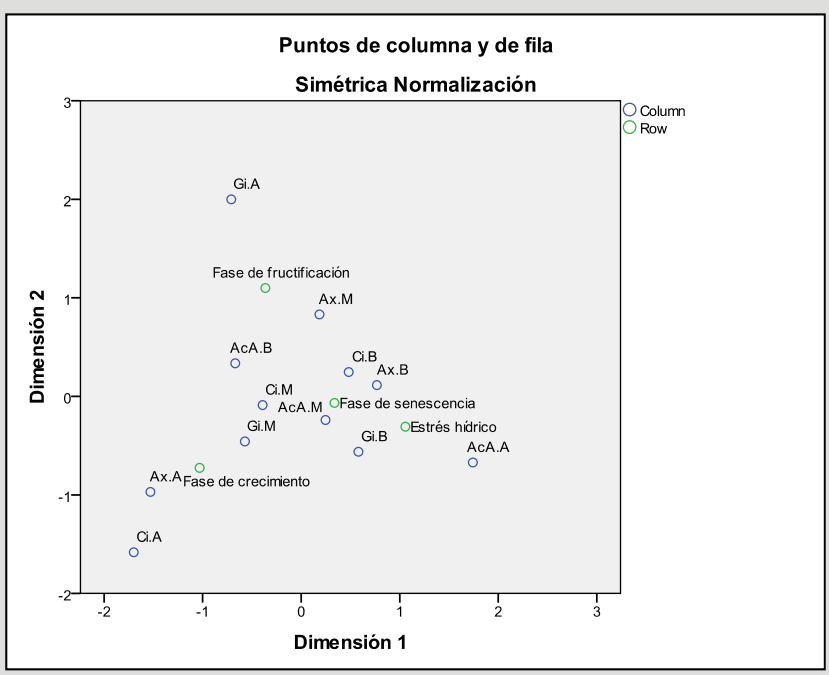
que debe acabar necesariamente con «.». Entre paréntesis se indica el número de filas (4) y el de columnas (12, sin tener en cuenta la columna «ROWCAT_») de nuestra tabla de frecuencias.

CUADRO XIII.3. *(Continuación)*[SPSS]


Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ejecutar Herramientas Ventana Ayuda
Activo: Conjunto_de_datos1
CORRESPONDENCE
1 CORRESPONDENCE TABLE=ALL(4,12).

Paso 8. A continuación seleccionamos «Ejecutar» ⇨ «Todo».

Paso 9. El gráfico muestra que el primer eje parece venir definido por el estrés hídrico y la fase en la cual la planta está creciendo. Concentraciones altas de ácido abscísico están asociadas al estrés hídrico y concentraciones altas de auxinas y citoquininas a la fase de crecimiento de la planta. El segundo eje viene determinado por la fase de fructificación, con valores altos de giberelina.



CUADRO XIII.3. *(Continuación)*[SPSS]

Paso 10. Los autovalores representan la varianza (o inercia) explicada por esas componentes: 54,45% la primera, 31,11% la segunda y 85,7% entre ambas.

Resumen

Dimensión	Valor propio	Inercia	Chi-cuadrado	Sig.	Proporción de inercia		Confianza para el Valor propio	
					Explicada	Acumulada	Desviación típica	Correlación 2
1	,607	,368			,545	,545	,042	,132
2	,458	,210			,311	,857	,064	
3	,311	,097			,143	1,000		
Total		,675	129,633	,000[a]	1,000	1,000		

a. 33 grados de libertad

Paso 11. Las tablas siguientes son para las filas y las columnas. Ambas tablas se interpretan del mismo modo. Observemos por ejemplo la tabla de las filas (las fases fisiológicas en las que se encuentra la planta). En primer lugar está la «Masa», que muestra la diferente importancia o peso de cada fase dentro del conjunto de datos, y todas tienen igual peso. Luego aparecen las coordenadas o «Puntuaciones», que permitirán la representación gráfica. La «Inercia» es la contribución absoluta o porcentaje de cada eje explicado por cada fase. Cuando la planta está en crecimiento (43,8%) y en estrés hídrico (46,1%) son las fases con mayor contribución a la primera componente, y la fase de fructificación (65,9%) a la segunda.

Examen de los puntos de fila[a]

Fila	Masa	Puntuación en la dimensión		Inercia	Contribución				
					De los puntos a la inercia de la dimensión		De la dimensión a la inercia del punto		
		1	2		1	2	1	2	Total
Estrés hídrico	,250	1,057	-,307	,203	,461	,052	,834	,053	,887
Fase de crecimiento	,250	-1,031	-,725	,224	,438	,287	,720	,269	,989
Fase de fructificación	,250	-,364	1,099	,162	,055	,659	,124	,854	,978
Fase de senescencia	,250	,338	-,067	,086	,047	,002	,202	,006	,208
Total activo	1,000			,675	1,000	1,000			

a. Normalización Simétrica

Para las hormonas, las que tienen más masa y, por tanto, más peso en el análisis son ácido abscísico en bajas concentraciones (13,5%), citoquininas en bajas concentraciones (13,5%) y giberelina a bajas concentraciones (13%). Las hormonas con más contribución a la inercia o varianza de la primera dimensión son auxinas a altas concentraciones (AX.A con 26,1%), y ácido abscísico a altas concentraciones (AcA.A con 20,8%).

CUADRO XIII.3. *(Continuación)*[SPSS]

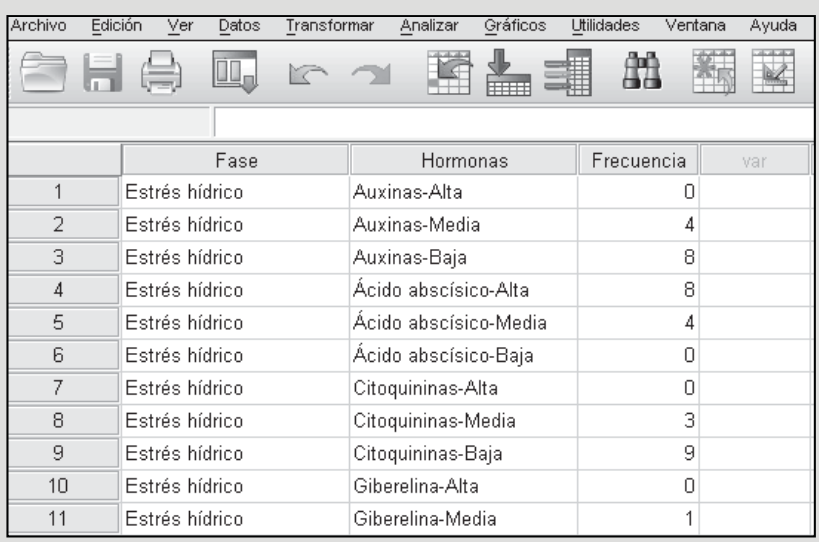
Examen de los puntos columna[a]

Columna	Masa	Puntuación en la dimensión 1	Puntuación en la dimensión 2	Inercia	Contribución: De los puntos a la inercia de la dimensión 1	Contribución: De los puntos a la inercia de la dimensión 2	Contribución: De la dimensión a la inercia del punto 1	Contribución: De la dimensión a la inercia del punto 2	Contribución: De la dimensión a la inercia del punto Total
Ax.A	,068	-1,530	-,970	,133	,261	,139	,725	,220	,945
Ax.M	,063	,186	,832	,031	,004	,094	,042	,634	,676
Ax.B	,120	,768	,114	,062	,116	,003	,688	,012	,700
AcA.A	,042	1,742	-,671	,125	,208	,041	,614	,069	,683
AcA.M	,073	,247	-,240	,007	,007	,009	,362	,259	,621
AcA.B	,135	-,669	,336	,049	,100	,033	,752	,143	,895
Ci.A	,016	-1,699	-1,582	,047	,074	,085	,584	,383	,966
Ci.M	,099	-,392	-,088	,010	,025	,002	,959	,037	,996
Ci.B	,135	,482	,247	,023	,052	,018	,822	,163	,985
Gi.A	,052	-,710	2,000	,119	,043	,454	,134	,804	,938
Gi.M	,068	-,571	-,458	,024	,036	,031	,567	,275	,842
Gi.B	,130	,581	-,562	,046	,072	,090	,584	,413	,998
Total activo	1,000			,675	1,000	1,000			

a. Normalización Simétrica

Por lo tanto, existe una clara asociación entre los niveles de hormonas y las fases fisiológicas de la planta. Cuando la planta está sometida a estrés hídrico hay altas concentraciones de ácido abscísico. Altas concentraciones de auxinas y citoquininas se producen cuando la planta crece. En la fase de fructificación se observan valores altos de giberelina. En la fase de senescencia, en general, hay valores bajos de la mayoría de las hormonas.

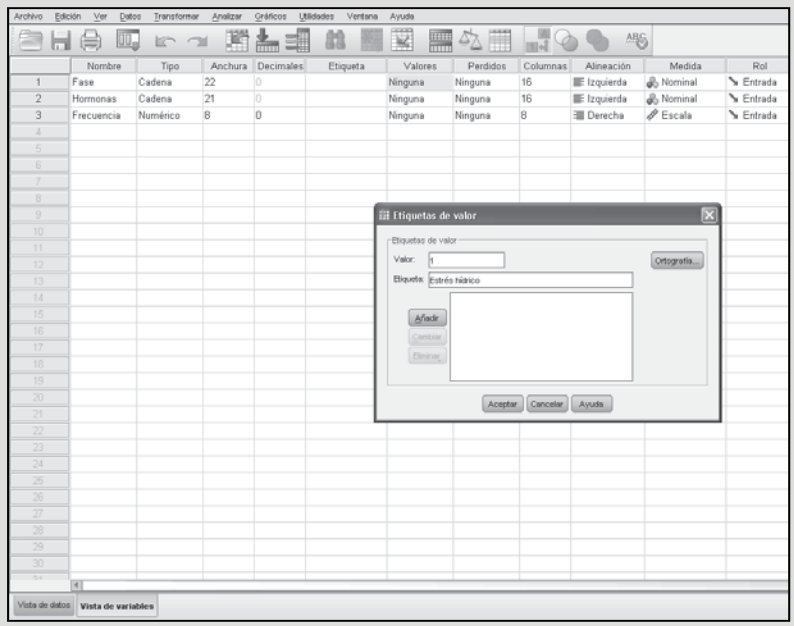
Paso 12. Si los datos están estructurados de la forma que se muestra en la siguiente ventana, el análisis se realiza de otra forma.



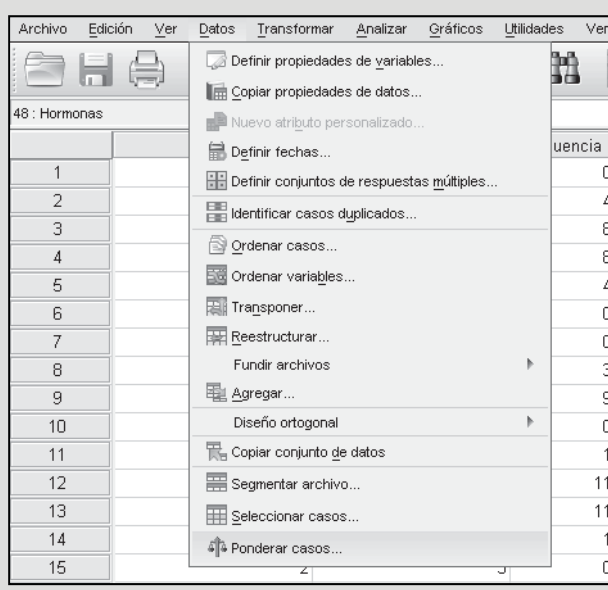
	Fase	Hormonas	Frecuencia	var
1	Estrés hídrico	Auxinas-Alta	0	
2	Estrés hídrico	Auxinas-Media	4	
3	Estrés hídrico	Auxinas-Baja	8	
4	Estrés hídrico	Ácido abscísico-Alta	8	
5	Estrés hídrico	Ácido abscísico-Media	4	
6	Estrés hídrico	Ácido abscísico-Baja	0	
7	Estrés hídrico	Citoquininas-Alta	0	
8	Estrés hídrico	Citoquininas-Media	3	
9	Estrés hídrico	Citoquininas-Baja	9	
10	Estrés hídrico	Giberelina-Alta	0	
11	Estrés hídrico	Giberelina-Media	1	

CUADRO XIII.3. *(Continuación)*[SPSS]

Paso 13. Es importante mencionar que los datos tienen que ser números y para ponerles etiquetas es necesario ir a la pestaña de «Valores» en la ventana «Vista de variables», como se muestra en la figura siguiente. Se debe especificar la etiqueta correspondiente a cada valor y pulsar «Añadir», tanto para las fases como para las hormonas.

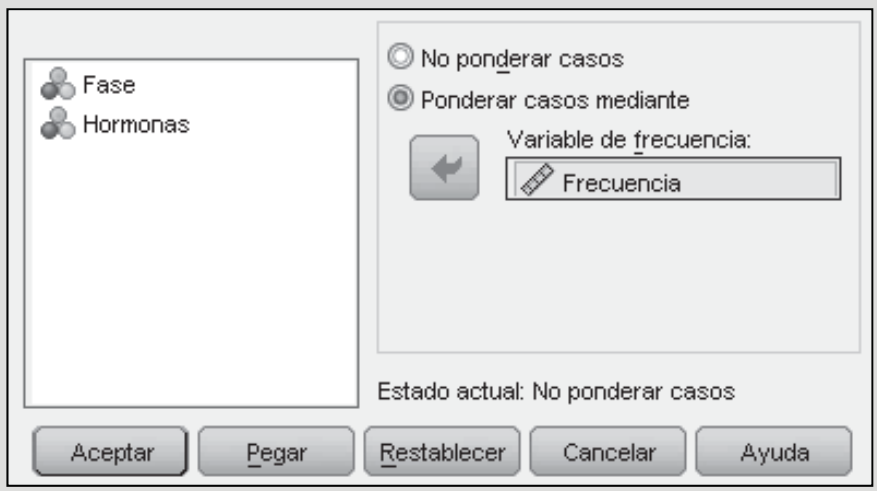

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana Ayuda
Nombre Tipo Anchura Decimales Etiqueta Valores Perdidos Columnas Alineación Medida Rol
1 Fase Cadena 22 0 Ninguna Ninguna 16 Izquierda Nominal Entrada
2 Hormonas Cadena 21 0 Ninguna Ninguna 16 Izquierda Nominal Entrada
3 Frecuencia Numérico 8 0 Ninguna Ninguna 8 Derecha Escala Entrada
Etiquetas de valor
Etiquetas de valor
Valor: 1
Etiqueta: Estrés hídrico
Ortografía...
Añadir
Cambiar
Eliminar
Aceptar Cancelar Ayuda
Vista de datos
Vista de variables

Paso 14. A continuación es necesario ponderar cada caso (fase y hormona) con su frecuencia correspondiente. Para ello seleccionamos «Datos» ⇨ «Ponderar casos…».

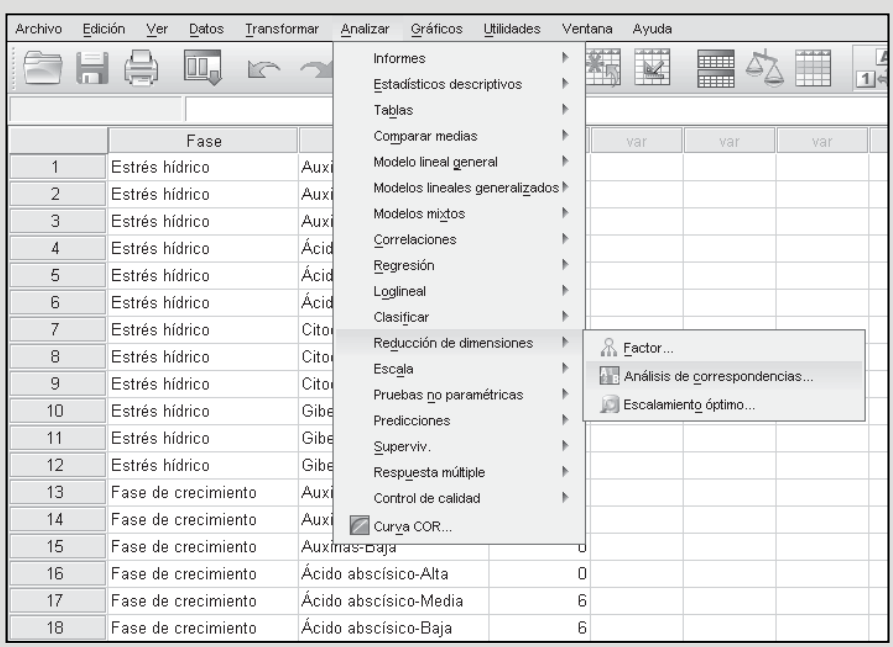


CUADRO XIII.3. *(Continuación)*[SPSS]

Paso 15. Se marca «Ponderar casos mediante» y en «Variable de frecuencia:» ponemos «Frecuencia».

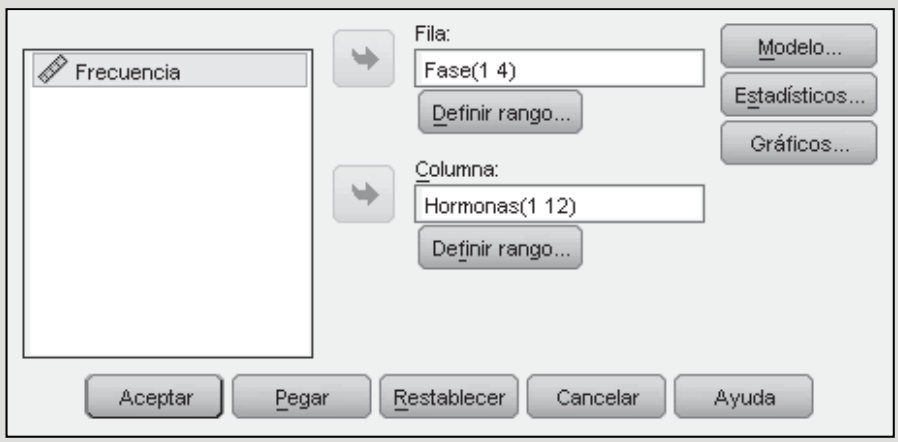


Paso 16. Ahora ya es posible realizar el análisis y, para ello, seleccionamos «Analizar» ⇨ «Reducción de dimensiones» ⇨ «Análisis de correspondencias…».



CUADRO XIII.3. *(Continuación)*[SPSS]

Paso 17. Por último, en «Fila» se introduce la variable «Fase» y en «Columna» la variable «Hormonas». En ambos casos hay que ir a «Definir rango…», y en la ventana que aparece introducir el rango de la variable «Fase» (de 1 a 4) y de la variable «Hormonas» (de 1 a 12).



Al pulsar «Aceptar» se obtienen los mismos resultados que se mostraron y explicaron anteriormente.

XIII.4. Escalamiento multidimensional

Trata de encontrar la estructura subyacente en un conjunto de medidas de proximidad/distancia entre objetos. Los objetos pueden ser de cualquier tipo: Países, marcas comerciales, relaciones de parentesco, empresas, descriptores de calidad, etc. Las medidas que permiten establecer distancias o proximidades entre objetos también pueden ser de distinto tipo, cuantitativas o cualitativas, reales o supuestas, objetivas o subjetivas. Se suele aplicar esta técnica al estudio de "percepciones", de gran interés en psicología, marketing, control de calidad, y otras áreas.

Si tenemos un mapa, es relativamente sencillo construir una tabla de distancias entre ciudades. El escalamiento multidimensional trata de resolver el problema inverso: a partir de una tabla de distancias se construye un mapa, que permite situar los objetos e interpretar gráficamente las relaciones existentes entre ellos.

Se asigna cada observación a una posición específica en un espacio de reducidas dimensiones, generalmente dos, de tal modo que las distancias entre los puntos se correspondan del mejor modo posible con las distancias entre los objetos. El resultado principal de esta técnica es por lo tanto un gráfico.

El criterio o medida de ajuste más importante utilizado en el escalamiento multidimensional es el Stress (Kruskal), que es un promedio de desviaciones entre las distancias finales en el mapa y las distancias o disimilaridades iniciales, normalizado para que tome valores entre 0 y 1.

$$S = \frac{\sum \sum \left(d_{ij} - \hat{d}_{ij} \right)^2}{\sum \sum \left(d_{ij}^2 \right)}$$

Cuanto más grande es el stress, peor es el ajuste, considerándose generalmente que un valor superior a 0,2 aconseja no utilizar esta técnica. A menudo se expresa en porcentaje: el ajuste es bueno cuando el stress es menor que el 5% (0,05). Para su interpretación se puede utilizar la siguiente escala orientativa:

Stress	**Ajuste**
0,1 - 0,2	Pobre
0,05 - 0,1	Regular
0,025 - 0,05	Bueno
<0,025	Excelente
0	Perfecto

También se emplea como medida de ajuste el coeficiente de correlación múltiple al cuadrado *RSQ* entre las distancias y las disparidades, que se puede interpretar como la proporción de varianza de los datos transformados -o porcentaje de variabilidad- recogida por la representación obtenida. Varía entre 0 y 1, y cuanto más próximo a 1 mejor es el ajuste; se suele considerar que 0,6 es un valor mínimo aceptable.

Existen numerosas variantes y algoritmos para la realización del Escalamiento Multidimensional.

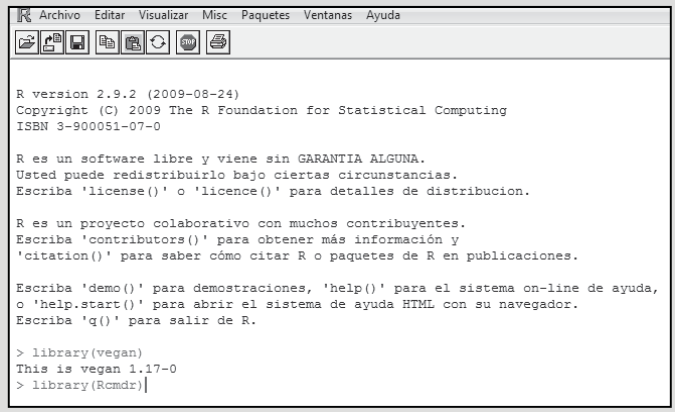
CUADRO XIII.4. Escalamiento multidimensional

Aplicación con R.

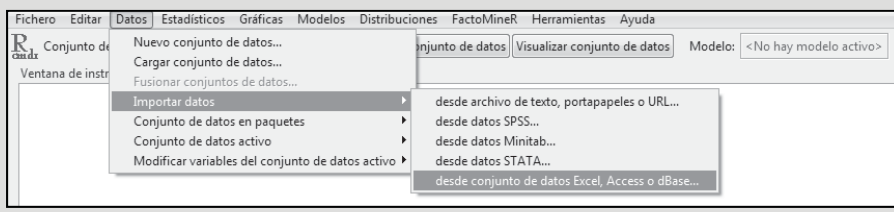
EJEMPLO. Se ha realizado una encuesta en la que se pide a un número elevado de consumidores que relacionen un conjunto de 10 pescados cocinados (lenguado, palometa, jurel, pez espada, lubina, rape, caballa, besugo, merluza, bacalao) con un conjunto de descriptores sensoriales (color, sabor, aroma, consistencia, aspecto, flexibilidad), otorgando puntuaciones de 0 a 5. Los datos del archivo **Cuadro XIII.4.xls** consisten en las puntuaciones medias obtenidas.

Paso 1. Iniciamos R y elegimos el directorio de trabajo. Son necesarios los paquetes "vegan" y "Rcmdr" y, de no estar instalados, la instalación se realiza como se explica en el Apéndice II. Se cargan o se incorporan con las instrucciones «library(vegan)» y luego «library(Rcmdr)».

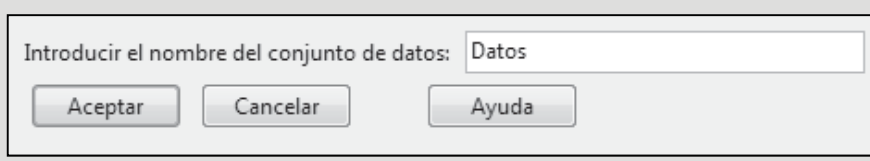
CUADRO XIII.4. *(Continuación)*[R]


Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> library(vegan)
This is vegan 1.17-0
> library(Rcmdr)

Paso 2. A continuación se selecciona «Datos» ⇨ «Importar datos» ⇨ «desde conjunto de datos Excel, Access o dBase».


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones FactoMineR Herramientas Ayuda
Nuevo conjunto de datos...
Cargar conjunto de datos...
Fusionar conjuntos de datos...
Importar datos
Conjunto de datos en paquetes
Conjunto de datos activo
Modificar variables del conjunto de datos activo
desde archivo de texto, portapapeles o URL...
desde datos SPSS...
desde datos Minitab...
desde datos STATA...
desde conjunto de datos Excel, Access o dBase...
Visualizar conjunto de datos
Modelo:
<No hay modelo activo>

Paso 3. Aparece un pequeño cuadro de diálogo para introducir el nombre del conjunto de datos, en el que dejamos el nombre por defecto «Datos» y pulsamos «Aceptar». Ahora aparecen los archivos en el directorio de trabajo, en el que elegimos **Cuadro XIII.4.xls** y pulsamos «Abrir».

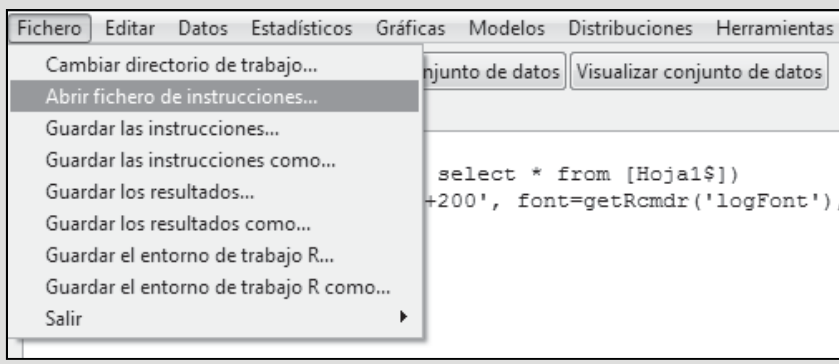


Paso 4. En R Commander pulsamos en «Visualizar conjunto de datos». Es importante inspeccionar los datos y comprobar que no existan líneas, lo cual a veces ocurre al final del archivo, que tengan celdas con el código <NA>. Si existen estas líneas, hay que eliminarlas.

	FILAID	Aspecto	Flexibilidad	Sabor	Color	Aroma	Consistencia
1	Lenguado	3.8133976	0.4389203	2.6288784	0.9447782	4.331075	1.5296606
2	Palometa	0.1246077	3.0018136	0.6477163	3.4151884	3.037923	0.5622651
3	Jurel	2.2734281	3.5514117	0.2937789	2.7759400	4.062515	1.0467295
4	Pez espada	0.6974218	1.0956345	0.4610865	0.6848375	3.301972	0.5140075
5	Lubina	3.6897341	3.4227827	4.2431039	3.0201022	4.655982	3.5391640
6	Rape	4.4237945	2.3307759	4.0495918	3.6464334	4.410111	1.7135144
7	Caballa	1.5543407	3.4687167	0.8648217	4.0286493	4.230400	0.5858823
8	Besugo	3.4256587	3.0688065	3.6861795	2.3623736	4.040060	1.7149480

CUADRO XIII.4. *(Continuación)*[R]

Paso 5. A continuación se carga el archivo de instrucciones **Cuadro XIII.4.R** seleccionando «Fichero» ⇨ «Abrir fichero de instrucciones…» y no se guarda la salida.



Paso 6. La primera de estas instrucciones realiza el escalamiento multidimensional y todos los resultados producidos se almacenan en la variable Datos.mds

```
Datos.mds <- metaMDS(Datos[,2:7])
```

La segunda construye el gráfico

```
plot(Datos.mds, type="n")
```

La tercera añade las etiquetas de fila

```
text(Datos.mds$points, labels = Datos[,1], col=2, cex=.7)
```

La cuarta las etiquetas de columna

```
text(Datos.mds$species, labels = colnames(Datos[,2:7]), col=3, cex=.7)
```

La quinta escribe las coordenadas de filas

```
Datos.mds$points
```

La sexta son las coordenadas de columnas

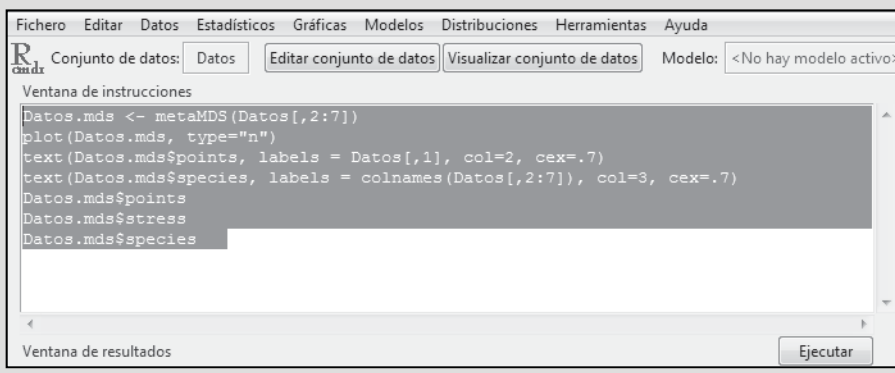
```
Datos.mds$species
```

y, por último, el valor del estadístico Stress

```
Datos.mds$stress
```

CUADRO XIII.4. *(Continuación)*[R]

Paso 7. Para realizar el análisis se selecciona todo y se pulsa «Ejecutar». Para cualquier otra aplicación con datos diferentes solo tenemos que modificar el rango de columnas (en el ejemplo, 2:7, entre la 2 y la 7) las dos veces que aparece, y asegurarnos de que en la primera columna del archivo con los datos están las etiquetas o identificadores de fila.



Paso 8. El estadístico de Stress es 3,14 %, lo cual indica un buen ajuste entre las distancias iniciales y las distancias resultantes en el gráfico y, por lo tanto, una alta fidelidad en la representación.

```
> Datos.mds$stress
[1] 3.144925
```

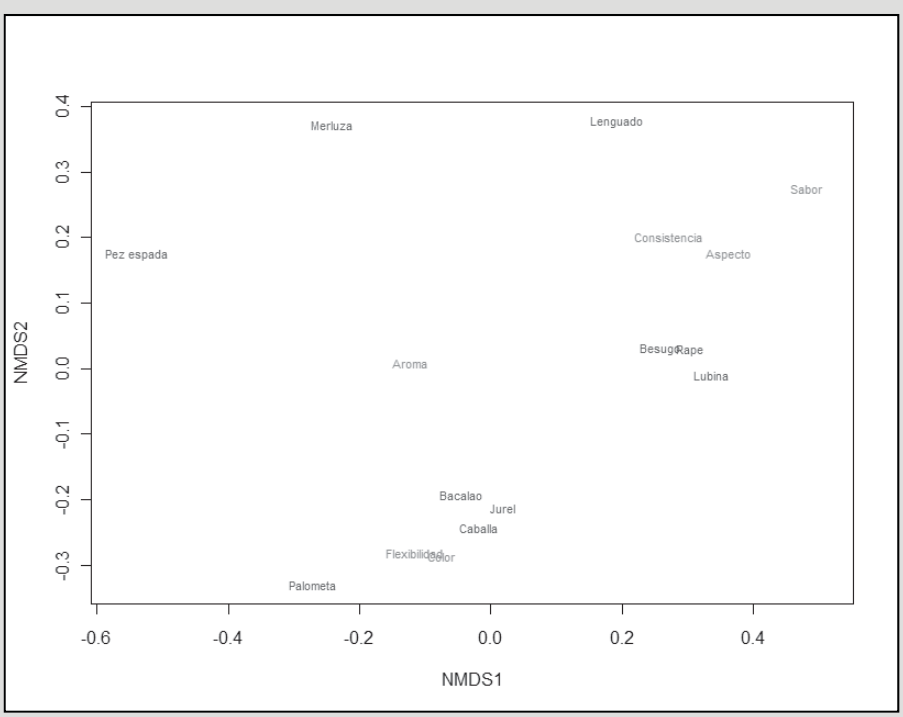
Paso 9. Las coordenadas han sido utilizadas para la representación gráfica.

```
> Datos.mds$points
           [,1]         [,2]
1   0.19201340  0.378286491
2  -0.26990583 -0.330170577
3   0.01902248 -0.212885823
4  -0.53830822  0.175407449
5   0.33691900 -0.008968118
6   0.30393447  0.029551648
7  -0.01723071 -0.242135303
8   0.25993453  0.030904919
9  -0.24189090  0.372008659
10 -0.04448820 -0.191999347
```

CUADRO XIII.4. *(Continuación)*[R]

```
> Datos.mds$species
                    [,1]         [,2]
Aspecto        0.36284449  0.175311820
Flexibilidad -0.11671162 -0.281626986
Sabor          0.48289863  0.274996602
Color         -0.07399728 -0.285728673
Aroma         -0.12335705  0.009213483
Consistencia   0.27055058  0.201479580
```

Paso 10. Lo más importante de este análisis es el gráfico, el cual muestra los diferentes tipos de peces y los descriptores de calidad. El hecho de que existan agrupaciones entre los pescados y los descriptores indica que los consumidores reconocen, entre las especies de peces, calidades diferentes. Sabor, aspecto y consistencia están relacionados con –próximos a– los pescados más apreciados (besugo, lenguado, lubina y rape), lo que indica que estos descriptores son los que parecen definir mejor –para los consumidores de nuestra muestra– la calidad. El descriptor aroma aparece en el centro, relacionado de modo similar –equidistante– con todos los pescados y, color y flexibilidad parecen relacionarse con los pescados menos apreciados.

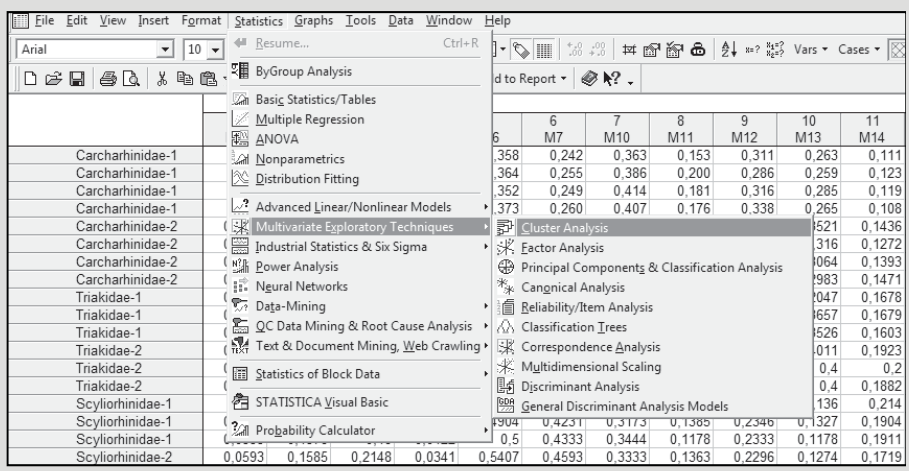


CUADRO XIII.4. *(Continuación)*[STATISTICA]

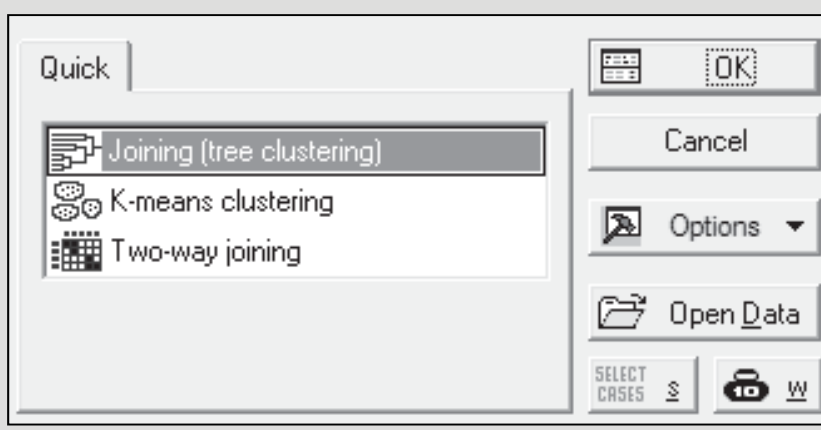
Aplicación con STATISTICA.

EJEMPLO. En el archivo **Cuadro XIII.4.sta** hay datos de medidas morfométricas de especies de tiburones, que pertenecen a distintas familias del orden Carcharhiniformes. El objetivo es determinar si las especies de la misma familia tienen morfometrías parecidas.

Paso 1. STATISTICA solo permite utilizar el escalamiento multidimensional cuando los datos tienen forma de matriz de distancias, con los mismos elementos en filas y columnas. Por lo tanto, en primer lugar debemos transformar nuestros datos brutos en una matriz de distancias. Se selecciona «Statistics» ⇨ «Multivariate Exploratory Techniques» ⇨ «Cluster Analysis».

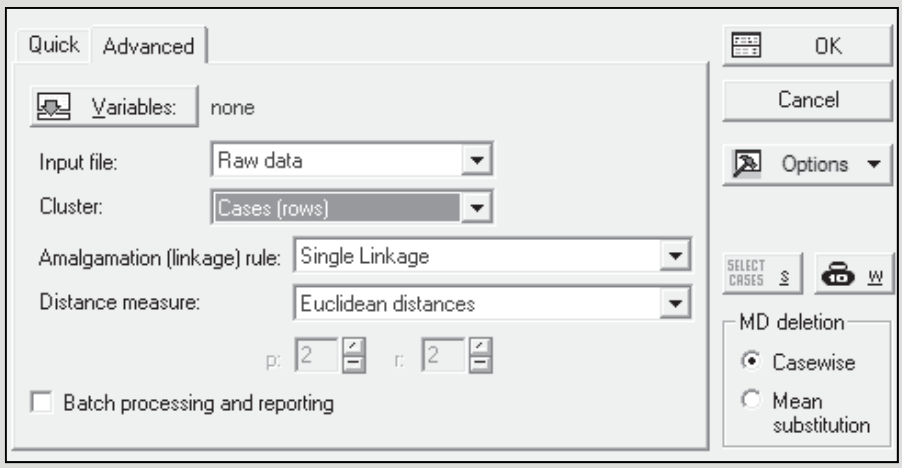


Paso 2. Se selecciona «Joining (tree clustering)».

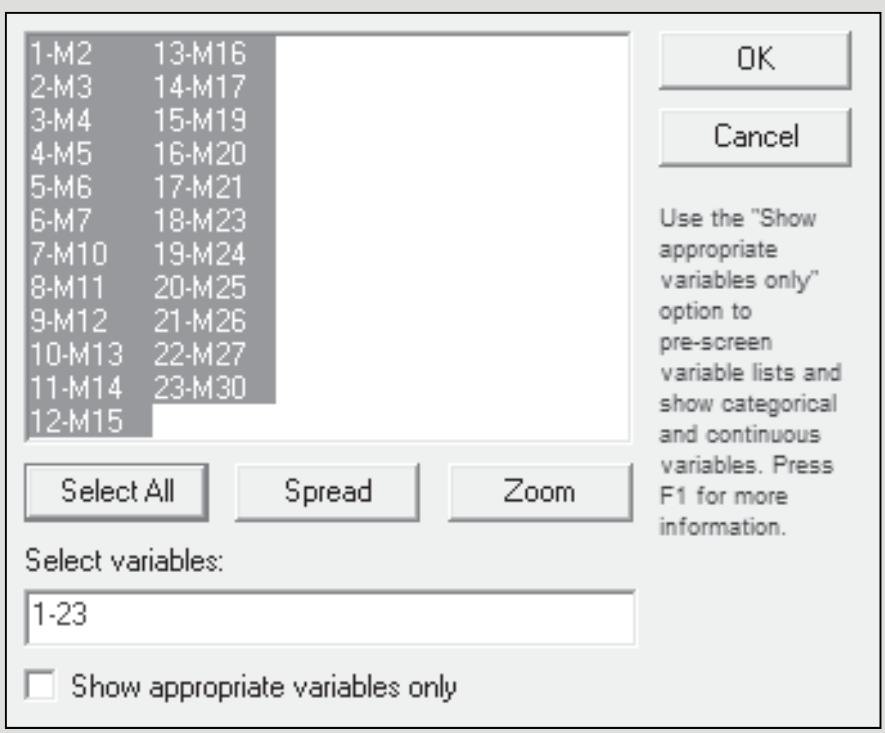


CUADRO XIII.4. *(Continuación)* *STATISTICA*

Paso 3. En la pestaña «Advanced» en «Cluster» se selecciona «Cases (rows)». Se pulsa en «Variables»

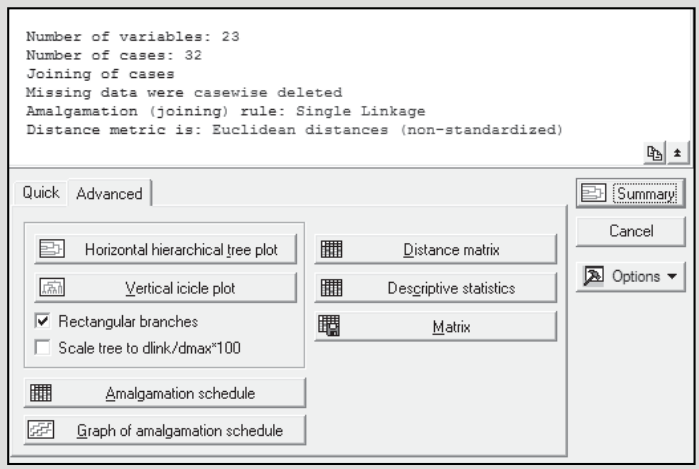


Paso 4. En la ventana que aparece se pulsa en «Select All».

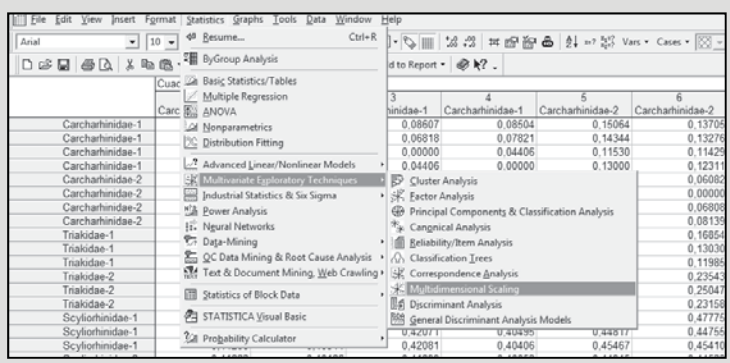


CUADRO XIII.4. *(Continuación)* STATISTICA

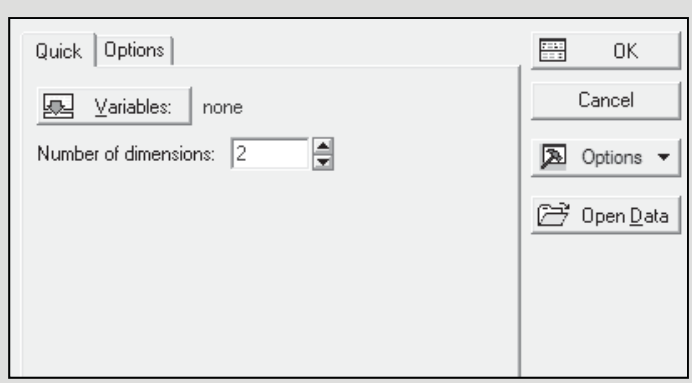
Paso 5. Al pulsar en «Matrix» se obtiene la matriz de distancias.



Paso 6. Ahora sobre la matriz de distancias es posible realizar el escalamiento multidimensional. Para ello se selecciona «Statistics» ⇨ «Multivariate Exploratory Techniques» ⇨ «Multidimensional Scaling».

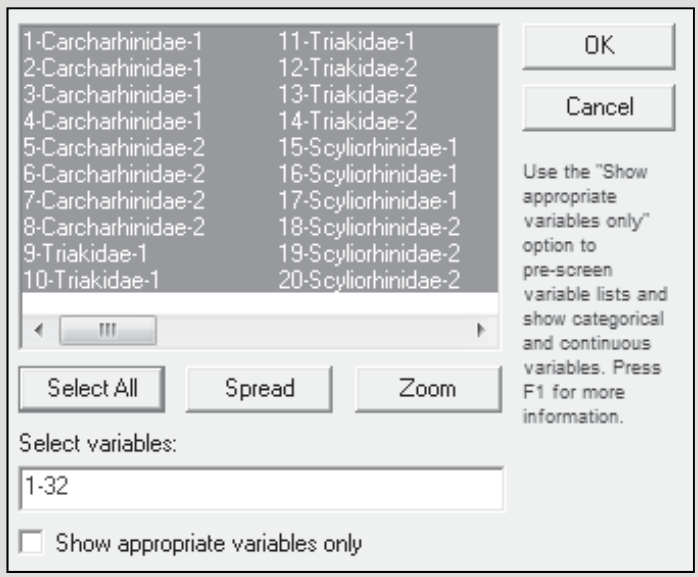


Paso 7. Se pulsa en «Variables».

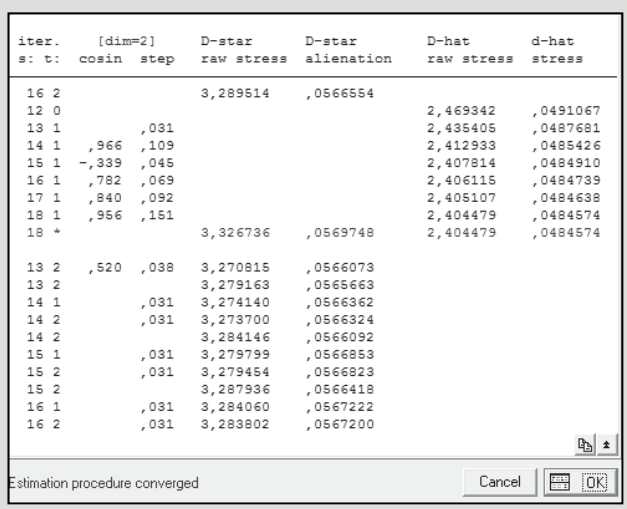


CUADRO XIII.4. *(Continuación)*[STATISTICA]

Paso 8. En la ventana que aparece se pulsa en «Select All».



Paso 9. Aparecen algunos resultados con los distintos pasos o iteraciones del algoritmo hasta alcanzar la convergencia. Debemos pulsar «Ok» para pasar a la ventana de resultados.

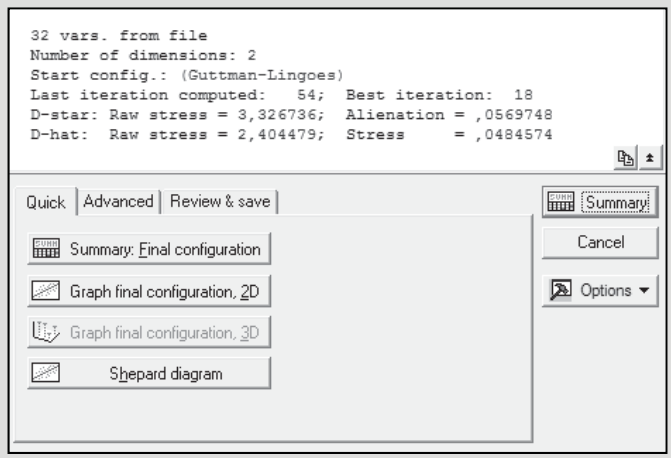


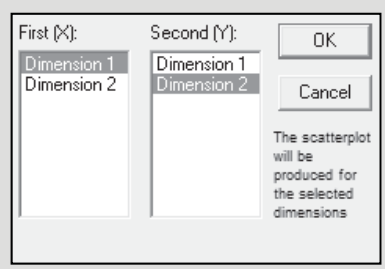
iters:	t:	[dim=2] cosin	[dim=2] step	D-star raw stress	D-star alienation	D-hat raw stress	d-hat stress
16	2			3,289514	,0566554		
12	0					2,469342	,0491067
13	1		,031			2,435405	,0487681
14	1	,966	,109			2,412933	,0485426
15	1	-,339	,045			2,407814	,0484910
16	1	,782	,069			2,406115	,0484739
17	1	,840	,092			2,405107	,0484638
18	1	,956	,151			2,404479	,0484574
18	*			3,326736	,0569748	2,404479	,0484574
13	2	,520	,038	3,270815	,0566073		
13	2			3,279163	,0565663		
14	1		,031	3,274140	,0566362		
14	2		,031	3,273700	,0566324		
14	2			3,284146	,0566092		
15	1		,031	3,279799	,0566853		
15	2		,031	3,279454	,0566823		
15	2			3,287936	,0566418		
16	1		,031	3,284060	,0567222		
16	2		,031	3,283802	,0567200		

Estimation procedure converged — Cancel — OK

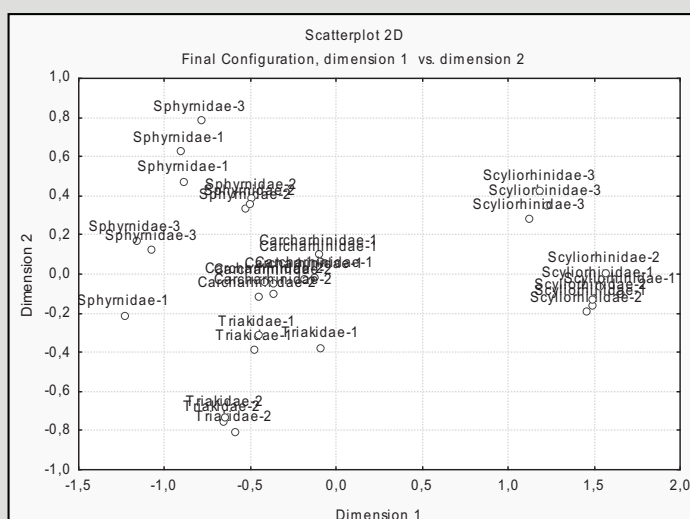
Paso 10. En la ventana de resultados se muestra el valor de la medida de ajuste, con un Stress de 0,048 (4,8%), que de acuerdo con la clasificación mostrada en la introducción se considera bueno. Pulsaremos en «Graph final configuration, 2D» para obtener el gráfico, principal resultado de este método; en la ventana que aparece se elige la «Dimension 1», en el eje *X*, frente a la «Dimension 2» en el eje *Y*.

CUADRO XIII.4. *(Continuación)* STATISTICA





Paso 11. Se observa que existe un agrupamiento relativamente bueno entre familias. Por tanto, se concluye que es posible diferenciar las distintas familias de tiburones en función de su morfometría. Sin embargo, no es posible ver cuáles son las variables morfométricas responsables de esa diferenciación entre familias.

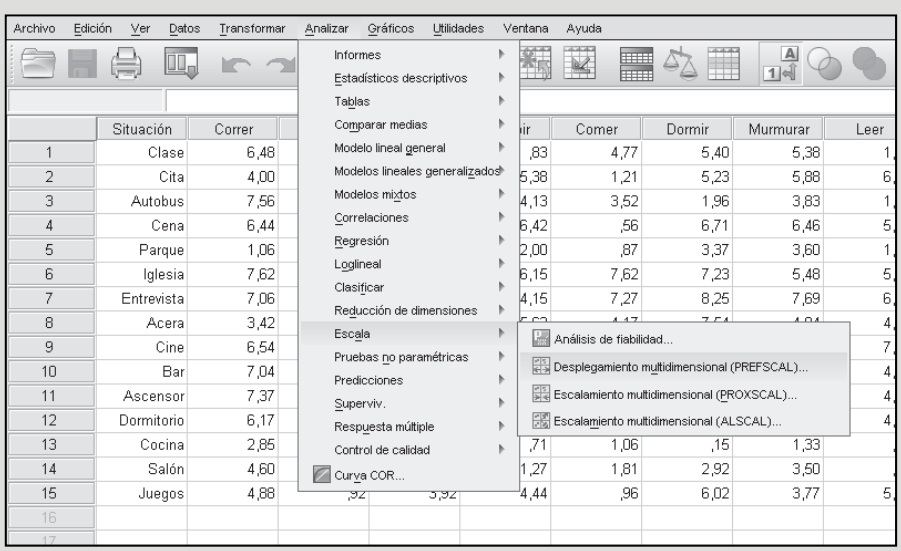


CUADRO XIII.4. *(Continuación)*[SPSS]

Aplicación con SPSS.

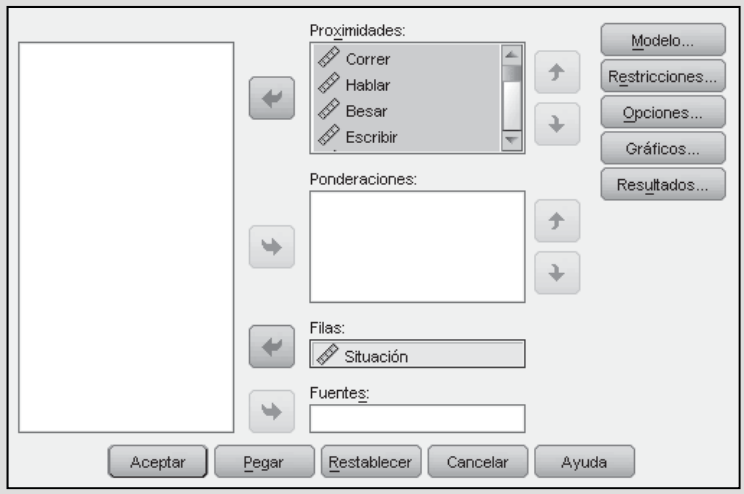
EJEMPLO. En un ejemplo clásico (Price & Bouffard, 1974), se pidió a 52 estudiantes que valoraran las combinaciones de 15 situaciones y 15 comportamientos en una escala de 10 puntos que oscila entre 0="totalmente apropiado" y 9="totalmente inapropiado". Los valores promediados respecto a los individuos se toman como disimilaridades. Los datos están en el archivo **Cuadro XIII.4.sav**. Los lugares o situaciones están en las filas y los comportamientos en las columnas. Se supone que conductas que pueden ser apropiadas en algunos lugares o situaciones no lo son en absoluto en otros.

Paso 1. Todas las variables deben ser numéricas, incluidas las etiquetas de los casos. Por tanto, si se quiere usar etiquetas en la variable que identifica las filas, es necesario realizarlo como se explicó en los pasos 1 a 3 del Cuadro IX.3. Existen distintas técnicas para el escalamiento multidimensional, de las cuales SPSS incorpora tres: dos de ellas aplicables a un único conjunto de objetos (que pueden ser casos o variables), que se diferencian fundamentalmente en que las medidas son distancias (ALSCAL) o proximidades (PROXSCAL), y la tercera aplicable a dos conjuntos diferentes de objetos –filas y columnas– denominada PREFSCAL, o desplegamiento multidimensional. Veremos un ejemplo de esta última, para lo cual se selecciona «Analizar» ⇨ «Escala» ⇨ «Desplegamiento multidimensional (PREFSCAL)…».

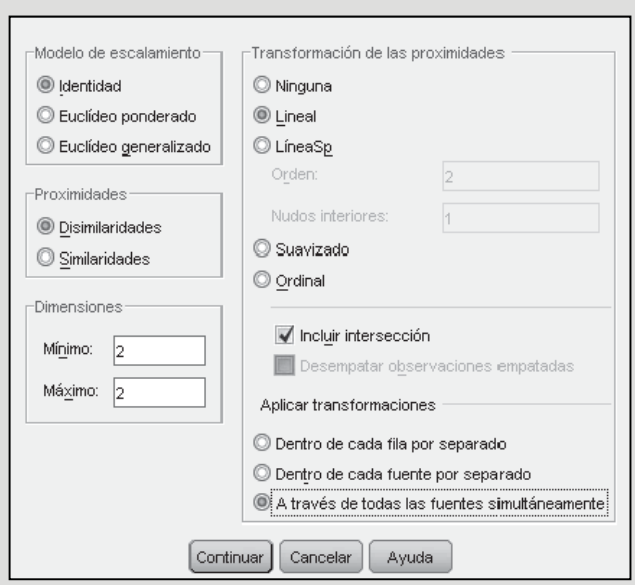


Paso 2. Se selecciona desde «Correr» hasta «Gritar» como variables de proximidad. Seleccionamos «Situación» como la variable de fila. Se pulsa en «Modelo».

CUADRO XIII.4. *(Continuación)*[SPSS]



Paso 3. Se selecciona «Lineal» como la transformación de proximidad, se marca «Incluir intersección», aplicar las transformaciones «A través de todas las fuentes simultáneamente» y en «Proximidades» se marca «Disimilaridades». La transformación lineal facilita el ajuste, y es adecuada cuando las diferencias entre puntuaciones son equivalentes en todo el rango de variación, es decir cuando en la escala utilizada un cambio en la puntuación de 1 a 2 es más o menos similar a un cambio de 6 a 7; esta transformación la aplicamos por igual a todos los datos. Marcamos «Disimilaridades» porque una puntuación más alta significa mayor inadecuación, es decir mayor distancia o disimilaridad entre los conceptos: «Escribir» es próximo a la situación «Clase», d = 0,83, pero distante de la situación «Cena», d = 6,42.



CUADRO XIII.4. *(Continuación)*[SPSS]

Paso 4. En el apartado «Medidas» de los resultados podemos observar algunos estadísticos de ajuste. En primer lugar el stress normalizado con un valor de 0,03586 (3,586%), que podemos considerar bueno. En segundo lugar el porcentaje de varianza explicada (80,45%) es también adecuado, ya que es bastante alto.

Medidas

Iteraciones		80
Valor de función final		,6427696
Partes del valor de función	Parte de Stress	,1893753
	Parte de penalización	2,1816613
Maldad de ajuste	Stress normalizado	,0358630
	Stress-I de Kruskal	,1893753
	Stress-II de Kruskal	,5214122
	S-Stress-I de Young	,2757464
	S-Stress-II de Young	,4526413
Bondad de ajuste	Dispersión explicada	,9641370
	Varianza explicada	,8084512
	Órdenes de preferencia recuperados	,8610714
	Rho de Spearman	,8983769
	Tau-b de Kendall	,7207219
Coeficientes de variación	Variación de las Proximidades	,5138436
	Variación de las Proximidades transformadas	,4737527
	Variación de las Distancias	,3906857
Índices de degeneración	Suma de cuadrados de los índices de entremezclado de DeSarbo	,4964590
	Índice de no-degeneración aproximada de Shepard	,7164286

Paso 5. Las tablas siguientes muestran las coordenadas de las situaciones y comportamientos, que se usan para representar la salida más importante del escalamiento multidimensional, el gráfico que se muestra a continuación.

Coordenadas de fila finales

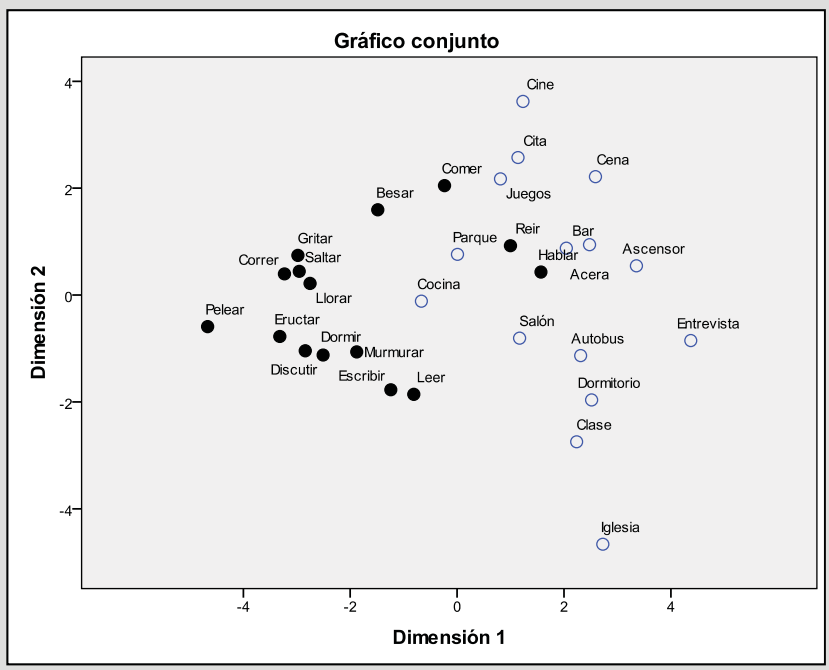
	Dimensión	
	1	2
Clase	2,237	-2,747
Cita	1,140	2,573
Autobus	2,312	-1,134
Cena	2,589	2,214
Parque	,008	,761
Iglesia	2,727	-4,663
Entrevista	4,373	-,852
Acera	2,479	,941
Cine	1,232	3,624
Bar	2,044	,875
Ascensor	3,355	,547
Dormitorio	2,516	-1,964
Cocina	-,668	-,114
Salón	1,170	-,807
Juegos	,811	2,174

Coordenadas de columna finales

	Dimensión	
	1	2
Correr	-3,230	,396
Hablar	1,569	,431
Besar	-1,484	1,594
Escribir	-1,240	-1,773
Comer	-,235	2,048
Dormir	-2,841	-1,042
Murmurar	-1,878	-1,064
Leer	-,810	-1,858
Pelear	-4,668	-,590
Eructar	-3,317	-,776
Discutir	-2,507	-1,122
Saltar	-2,953	,444
Llorar	-2,752	,217
Reir	,998	,923
Gritar	-2,979	,742

CUADRO XIII.4. *(Continuación)*[SPSS]

Paso 6. En algunos casos los puntos aparecen completamente agrupados en el centro del gráfico por un defecto de escala y el gráfico no es legible. Si es así debemos editar el gráfico (con doble clic sobre él), y ya dentro del editor de gráficos pulsaremos el icono *X* (debajo de editar), para marcar las casillas «Automático» en el mínimo y máximo. Haremos lo mismo con *Y* para ajustar la escala del eje vertical.



La dimensión horizontal parece estar más asociada a los objetos de columna (comportamientos) y discrimina entre comportamientos menos apropiados (pelear, eructar, discutir, correr) y comportamientos más apropiados (leer, comer, hablar, reír). Cuanto más hacia la izquierda, menos apropiado.

La dimensión vertical parece estar más asociada a los objetos de fila (situaciones) y define distintas restricciones situacionales o de comportamiento. Hacia la parte inferior hay situaciones (iglesia, clase) que restringen los comportamientos a aquellos menos conflictivos, y más introspectivos (leer, escribir). Hacia la parte superior están las situaciones (cine, juego, cita) que permiten los tipos de comportamiento más sociales o extrovertidos (comer, besar, reír).

En el centro las situaciones se separan en la dimensión horizontal según lo restrictivo que resulta en general la situación. Las más alejadas de los comportamientos (entrevista) son las más restrictivas, mientras que las más cercanas a los comportamientos (cocina, parque) son las menos restrictivas.

Capítulo XIV

ANÁLISIS MULTIVARIANTE: MÉTODOS DE CLASIFICACIÓN

Es el conjunto de métodos de análisis multivariante cuyo objetivo principal es agrupar o clasificar a los elementos de una muestra de objetos descritos mediante varias variables. Frente a los métodos factoriales, las técnicas de clasificación tienen en general un planteamiento algorítmico (no algebraico), y la mayoría de ellas no están basadas en el modelo lineal. Históricamente surgen con el computador y la informática, y se benefician de la extraordinaria potencia de cálculo que proporcionan estas herramientas. Existen centenares de métodos de clasificación o análisis cluster, de los cuales veremos solamente los más importantes.

En general se trata de obtener grupos (clases, clusters) homogéneos, es decir construidos de tal forma que los elementos de cada grupo sean similares entre sí; las clases deben ser diferentes o separadas, de modo que sean muy distintos los elementos de grupos diferentes. Se busca una estructura en los datos que ayude a comprender mejor la realidad de la población investigada.

Muchos métodos de clasificación permiten utilizar variables de tipo diverso, cuantitativas o cualitativas, ordinales o nominales, y en general no se necesitan las hipótesis restrictivas, como Normalidad y homogeneidad de varianzas, habituales en otras técnicas estadísticas. Tampoco es necesario en general un conocimiento previo de los datos.

Por estas razones los métodos de clasificación se suelen emplear con fines exploratorios, para establecer una primera aproximación a los datos, y para generar ideas, plantear hipótesis que puedan ser contrastadas con otras técnicas, o para iniciar la construcción de modelos explicativos.

XIV.1. Clasificación jerárquica

Es uno de los métodos más utilizados. Se obtiene una secuencia de particiones (clasificaciones) organizada jerárquicamente, desde una única clase que contiene a todos los elementos, hasta *n* clases, cada una de ellas con un solo elemento. La secuencia puede ser representada gráficamente mediante un árbol.

El algoritmo consiste en unir en primer lugar los dos elementos más próximos entre sí; la clase resultante se considera un nuevo elemento que sustituye a los dos anteriores. A continuación se buscan de nuevo los dos elementos más próximos entre sí, que se unirán en una clase nueva. El proceso continúa, y las diferentes clases se van uniendo sucesivamente en clases más grandes, hasta que todos los elementos forman finalmente una sola clase.

XIV.1.1. CRITERIO DE DISTANCIA

Para decidir cuáles son los elementos de la muestra más próximos entre sí es necesario disponer de un criterio de distancia. Existen diferentes definiciones o tipos de distancia y es preciso elegir la más adecuada para los datos utilizados.

- *Euclídea*: Es la más utilizada, y se corresponde con la distancia geométrica habitual en un espacio multidimensional. Se define como la raíz cuadrada de la suma de los cuadrados de las diferencias en los valores de cada variable para los dos elementos considerados. La mayoría de los programas estadísticos consideran como opción por defecto la distancia euclídea.
- *Distancia Euclídea al cuadrado*: Tiene como efecto asignar un mayor peso a los objetos más alejados.
- *Distancia Manhattan*: Es el promedio de la diferencia entre dimensiones. El nombre proviene de su relación con la distancia real entre dos puntos en una ciudad con calles que se cortan perpendicularmente, donde en general la línea recta no es un camino posible.
- *Distancia de Mahalanobis*: Cada dimensión está ponderada por la inversa de su desviación típica. Esta distancia tiene una importante propiedad: es invariante ante cambios en la unidad de medida, lo que la hace especialmente útil cuando las variables son heterogéneas.

XIV.1.2. CRITERIO DE AGLOMERACIÓN

Cuando comenzamos el proceso de clasificación, al unir en una sola clase los dos elementos más próximos entre sí, esta clase pasa a considerarse un nuevo elemento que sustituye a los dos antiguos. Para continuar el algoritmo es necesario ahora calcular la distancia entre ese nuevo elemento y todos los anteriores, y para ello existen distintas opciones, que determinarán el criterio de aglomeración o conglomeración.

Existen diferentes criterios de distancia entre clases, o entre clases y elementos, que se corresponden con el concepto matemático de ultramétrica, más restrictivo que el concepto clásico de distancia o métrica:

- *Enlace simple* (vecino más próximo): La distancia entre dos grupos es la distancia entre los dos objetos más próximos entre los diferentes grupos. Emplear este procedimiento de conglomeración puede dar lugar en ocasiones a un "efecto de encadenamiento", por el cual en cada fase se une una clase con un elemento aislado en lugar de unirse entre sí dos clases con varios elementos cada una, dando lugar a una secuencia jerárquica de clasificaciones atípicas de escasa utilidad práctica.
- *Enlace completo* (vecino más lejano): La distancia entre dos grupos es la mayor distancia entre dos objetos de ambos grupos.
- *Enlace medio* o vinculación inter-grupo: distancia media entre todos los pares de objetos de los dos grupos.
- *Vinculación intra-grupo*: distancia media entre todos los pares de objetos del grupo unión de ambos grupos. Considera en el promedio no solo las distancias entre elementos de los dos grupos, sino también las distancias entre elementos del mismo grupo.

- *Método de McQuitty*: la distancia a cualquier objeto desde un grupo que es unión de otros dos es la media de las distancias de aquellos dos ponderada por el número de unidades de cada grupo.
- *Enlace de centroides*: distancia entre las medias o centroides de los dos grupos.
- *Método de Ward*: En cada paso se unen los dos elementos que dan lugar a una menor pérdida de información, considerada como la suma de cuadrados de distancias de cada objeto al centro de su clase (mínima varianza o inercia).

XIV.1.3. DENDROGRAMA

El principal resultado de esta técnica es un gráfico, denominado dendrograma o árbol de clasificación, que permite ver el proceso completo de clasificación desde su inicio con cada elemento en una clase hasta el final con una única clase en la que están todos los elementos de la muestra. En el gráfico se ven todas las particiones o clasificaciones obtenidas y su relación jerárquica. Para su mejor interpretación se pueden etiquetar los casos o elementos clasificados.

A partir del dendrograma es posible elegir la partición de mayor interés, en función de las distancias entre clases, que se miden en la escala lateral. La distancia entre dos clases viene dada por la longitud de la rama que las une en el árbol; en general se busca en el árbol la mayor separación entre particiones, que permite visualizar la clasificación más natural del conjunto de objetos.

El algoritmo exige en cada paso el cálculo de la matriz de distancias entre todos los elementos; cuando la muestra de datos contiene un número grande de objetos (varios cientos, o más de un millar) el procedimiento se hace excesivamente lento, de tal modo que puede requerir muchos minutos o incluso horas antes de alcanzar el resultado final. Por otra parte el gráfico obtenido, dendrograma, no es fácilmente interpretable cuando es demasiado grande. El método de clasificación jerárquica no es útil para muestras grandes, aunque en ese caso puede todavía ser aplicado a una parte de los datos (a una muestra aleatoria de los datos).

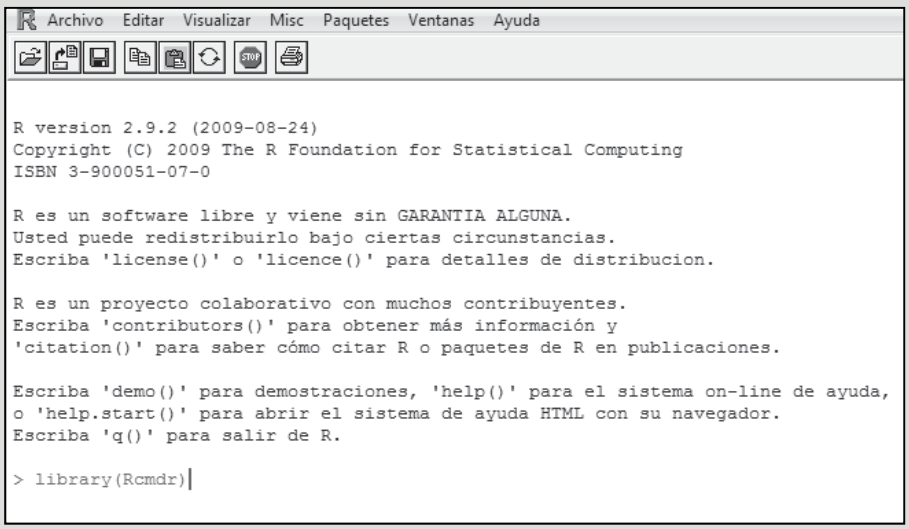
CUADRO XIV.1. Clasificación jerárquica

Aplicación con R.

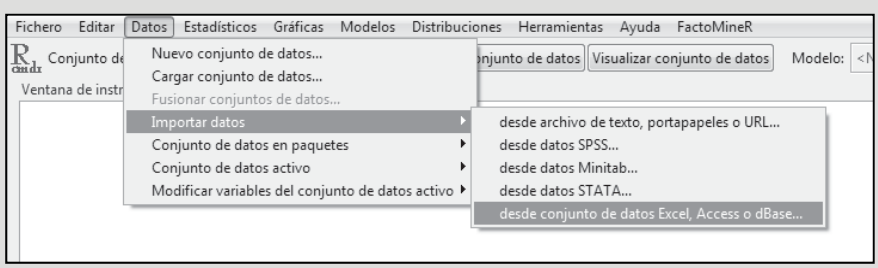
EJEMPLO. Tenemos una muestra de diferentes vinos de Galicia. Algunos han sido elaborados con la variedad de uva Albariño (A) y otros con la variedad Godello (G) y hay varios vinos que se desconoce a qué variedad pertenecen (D). Se han determinado las concentraciones de diferentes ácidos orgánicos (galacturónico, tartárico, málico, shiquímico, cítrico, succínico). El objetivo es averiguar si nuestros vinos se agrupan en clases razonablemente homogéneas, que pudieran corresponderse con la variedad de uva utilizada, o quizás con otras características determinables, como la zona de procedencia o el año de la cosecha. Los datos están en el archivo **Cuadro XIV.1.xls**.

CUADRO XIV.1. *(Continuación)*[R]

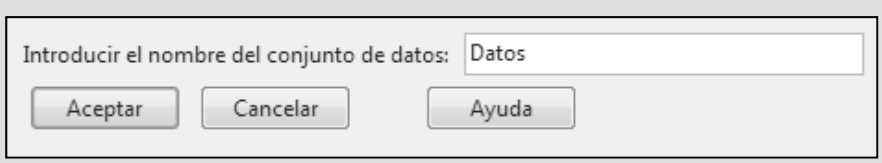
Paso 1. Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, cargamos el paquete "Rcmdr" con la instrucción library(Rcmdr). Si el paquete no está instalado, la instalación se realiza como se explica en el Apéndice II.


Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> library(Rcmdr)

Paso 2. Se selecciona «Datos» ⇨ «Importar datos» ⇨ «desde conjunto de datos Excel, Access o dBase».


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda FactoMineR
Nuevo conjunto de datos...
Cargar conjunto de datos...
Fusionar conjuntos de datos...
Importar datos
Conjunto de datos en paquetes
Conjunto de datos activo
Modificar variables del conjunto de datos activo
desde archivo de texto, portapapeles o URL...
desde datos SPSS...
desde datos Minitab...
desde datos STATA...
desde conjunto de datos Excel, Access o dBase...
Visualizar conjunto de datos
Modelo:

Paso 3. Aparece un pequeño cuadro de diálogo para introducir el nombre del conjunto de datos, en el que dejamos el nombre por defecto «Datos» y pulsamos «Aceptar». Posteriormente aparecen los archivos del directorio de trabajo, en el que elegimos **Cuadro XIV.1.xls** y pulsamos «Abrir».

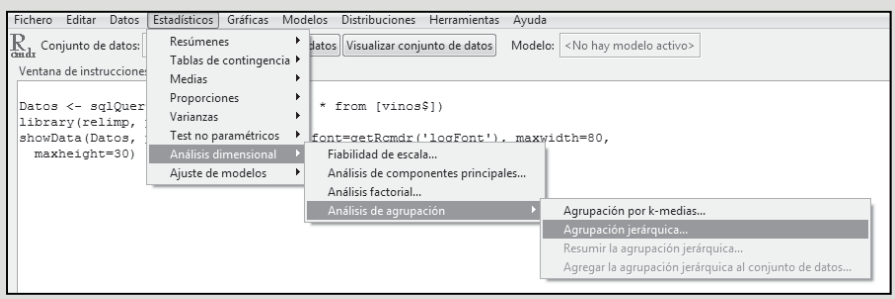


CUADRO XIV.1. *(Continuación)*[R]

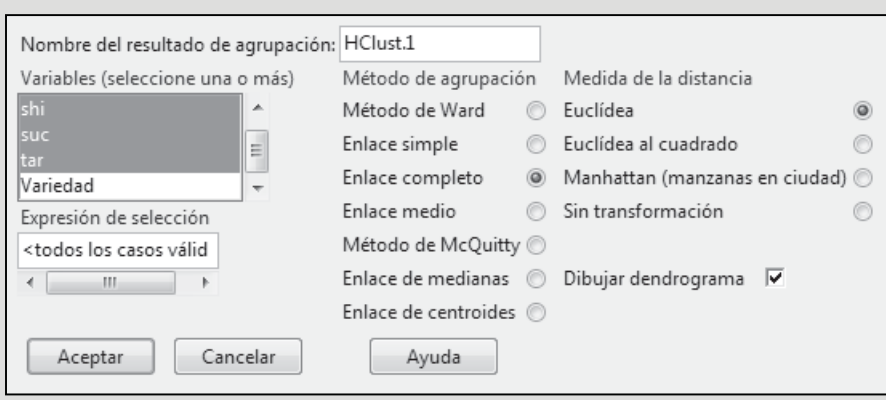
Paso 4. En R Commander pulsamos en «Visualizar conjunto de datos». Es importante inspeccionar los datos y comprobar que no existan líneas vacías, que contienen exclusivamente el código <NA>. Si existen estas líneas, hay que eliminarlas.

	Variedad	Etiqueta	gal	tar	mal	shi	cit	suc
1	2	G	0.482	1.796	2.562	0.006	0.136	0.236
2	1	A	0.429	2.749	3.458	0.032	0.138	0.170
3	2	G	0.453	2.193	2.716	0.007	0.152	0.311
4	2	G	0.375	2.048	3.055	0.007	0.094	0.268

Paso 5. Para aplicar el método de clasificación jerárquica elegimos en el menú principal de R Commander las opciones «Estadísticos» ⇨ «Análisis dimensional» ⇨ «Análisis de agrupación» ⇨ «Agrupación jerárquica…».

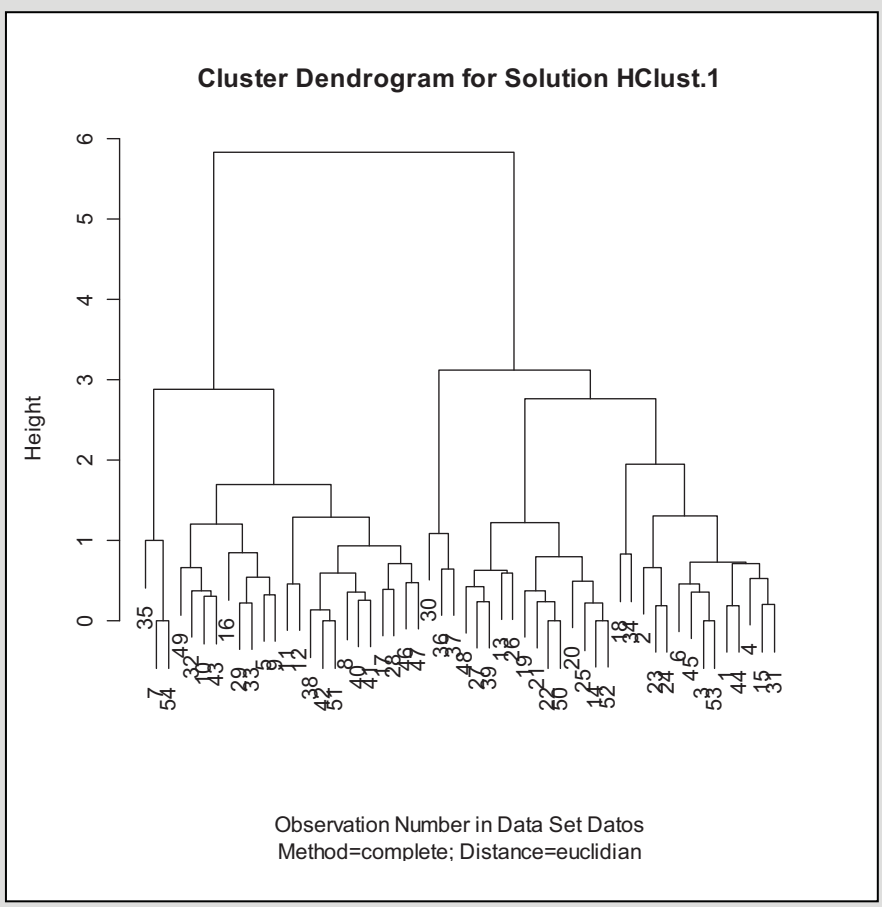


Paso 6. En la ventana de Agrupación jerárquica seleccionamos solamente las variables cuantitativas, desde «CIT» hasta «TAR», mantenemos el nombre por defecto asignado al resultado de agrupación «HClust.1», elegimos como método de agrupación «Enlace completo», como medida de distancia «Euclídea» y marcamos «Dibujar dendrograma».



CUADRO XIV.1. *(Continuación)*[R]

Paso 7. Estudiamos el dendrograma resultante. En sentido horizontal están los elementos de la muestra, identificados por su número de orden. En sentido vertical se muestran las distancias entre clases. Si cortamos con una línea horizontal a diferentes alturas podemos obtener clasificaciones con distinto número de grupos: 2 si el corte está entre 3 y 6 en la escala lateral, 5 si cortamos aproximadamente entre 2 y 3. Estas dos clasificaciones, especialmente la primera, son las que aparecen con más claridad. Para obtener otro número de grupos es necesario establecer la altura con más precisión, pero podemos encontrar cualquier número de clases entre 1 (arriba de todo) y 54 -el tamaño de muestra- (abajo de todo).



La clasificación de mayor interés es en general la que tiene un mayor desplazamiento en la escala lateral, en nuestro ejemplo dos clases. Vemos como estas dos clases se subdividen a su vez en otras, con diferentes clasificaciones organizadas jerárquicamente.

CUADRO XIV.1. *(Continuación)*[R]

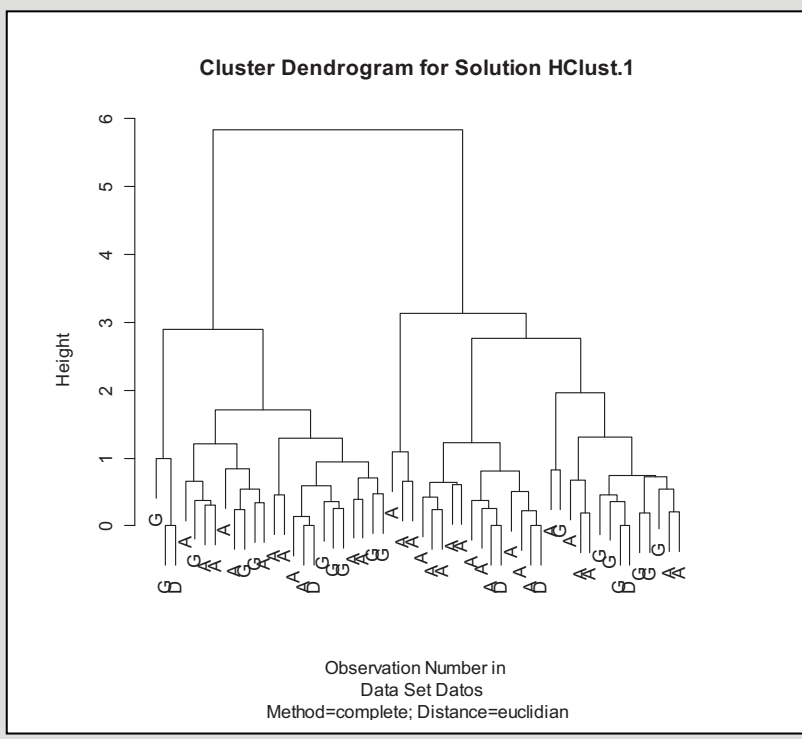
Paso 8. Podemos añadir al gráfico etiquetas que permitan una mejor comprensión de los objetos representados. Para ello modificamos en la ventana de instrucciones de R Commander la orden:

plot(HClust.1, main= "Cluster Dendrogram for Solution HClust.1", xlab= "Observation Number in Data Set Datos", sub="Method=complete; Distance=euclidian")

por:

plot(HClust.1, **label= Datos[,2],** main = "Cluster Dendrogram for Solution HClust.1", xlab= "Observation Number in Data Set Datos", sub="Method=complete; Distance=euclidian")

Es decir, sólo se introduce lo que está resaltado en negrita. «Datos» es el nombre que R commander asigna por defecto a nuestra tabla y en la columna 2 están las etiquetas que queremos utilizar para cada fila. Para ejecutar la orden, la seleccionamos con el cursor y pulsamos «Ejecutar».



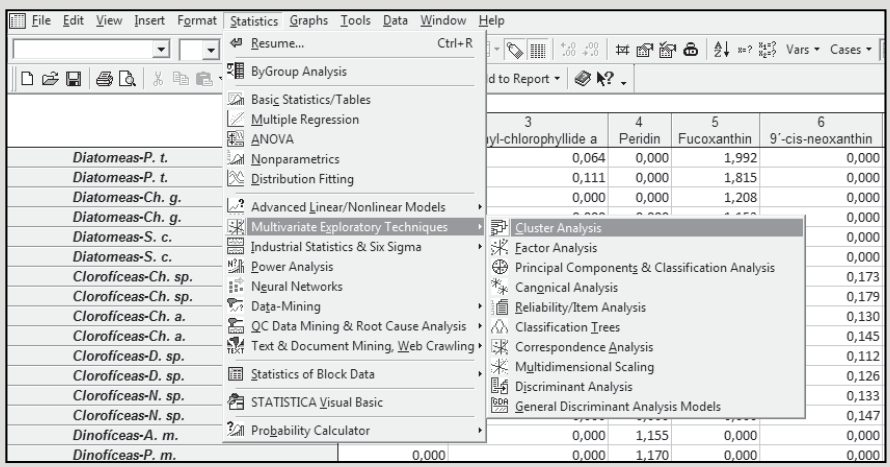
La etiqueta muestra una *A* cuando la variedad de uva es Albariño, *G* cuando es Godello y *D* cuando es desconocida. Ahora se observa que la gran clase de la derecha se subdivide en tres clases, de las cuales dos contienen vinos en los que no está la variedad Godello. Por lo tanto, la variedad de uva (Albariño y Godello) no es el criterio principal por el que se forman las dos grandes clases, pero tiene una cierta influencia en la clasificación.

CUADRO XIV.1. *(Continuación)*[STATISTICA]

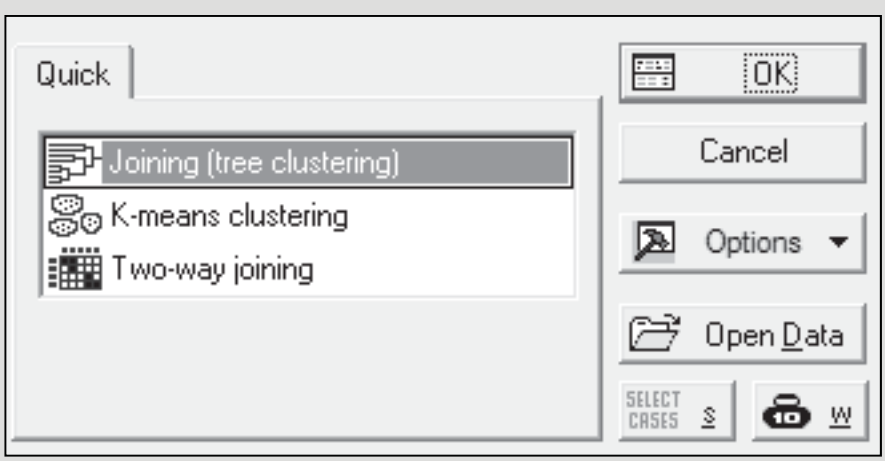
Aplicación con STATISTICA.

EJEMPLO. En el archivo **Cuadro XIV.1.sta** hay datos de la concentración relativa de 19 pigmentos, en relación con la concentración de clorofila *a*, de especies del fitoplancton que pertenecen a diferentes clases de algas: Diatomeas, Clorofíceas, Dinofíceas, Eustigmatofíceas, Cianofíceas y Criptofíceas. El objetivo es determinar si es posible diferenciar las clases de algas en función de su composición relativa de pigmentos.

Paso 1. Se selecciona «Statistics» ⇨ «Multivariate Exploratory Techniques» ⇨ «Cluster Analysis».

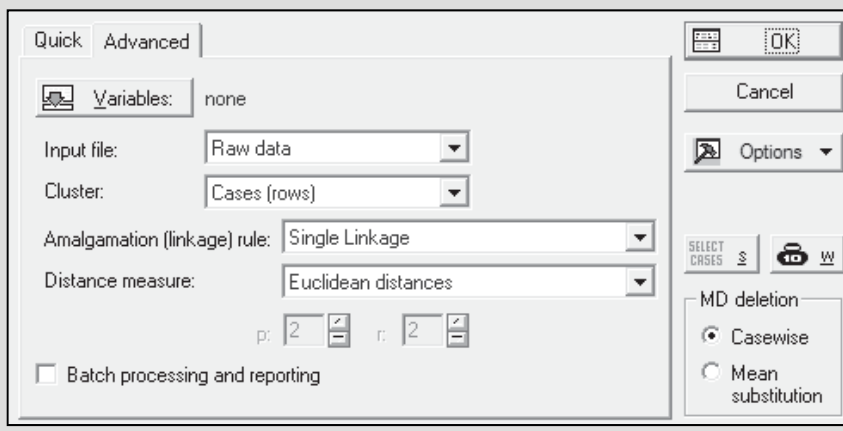


Paso 2. A continuación «Joining (tree clustering)».

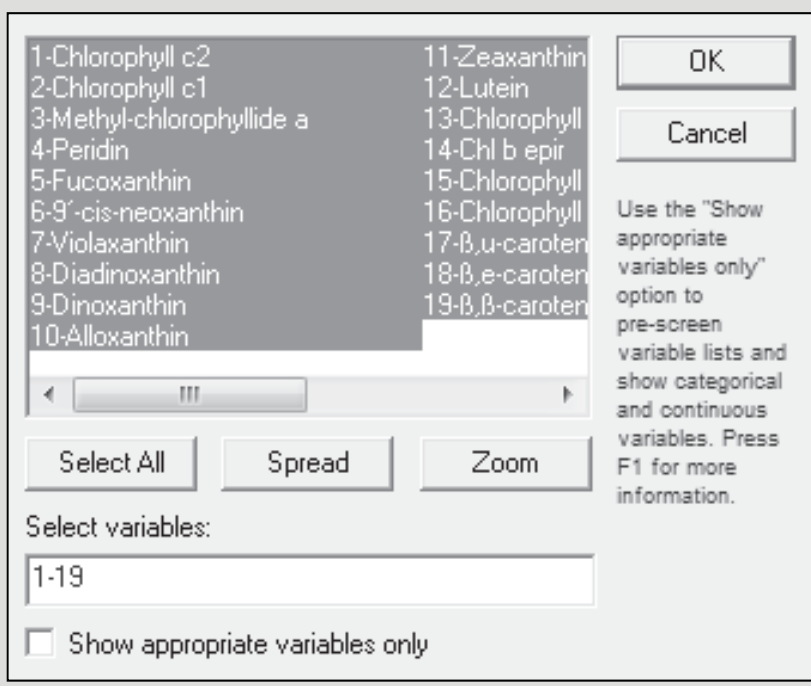


CUADRO XIV.1. *(Continuación)*[STATISTICA]

Paso 3. En la ventana que aparece pulsamos en la pestaña «Advanced», y mantenemos las opciones por defecto en la ventana de clasificación jerárquica: Datos brutos «Raw data», ya que nuestros datos no son una matriz de distancias, enlace simple o vecino más próximo «Single linkage», y «Ecuclidean distances». En la ventana «Cluster» elegiremos clasificar los casos, «Cases (rows)» en lugar de las variables.

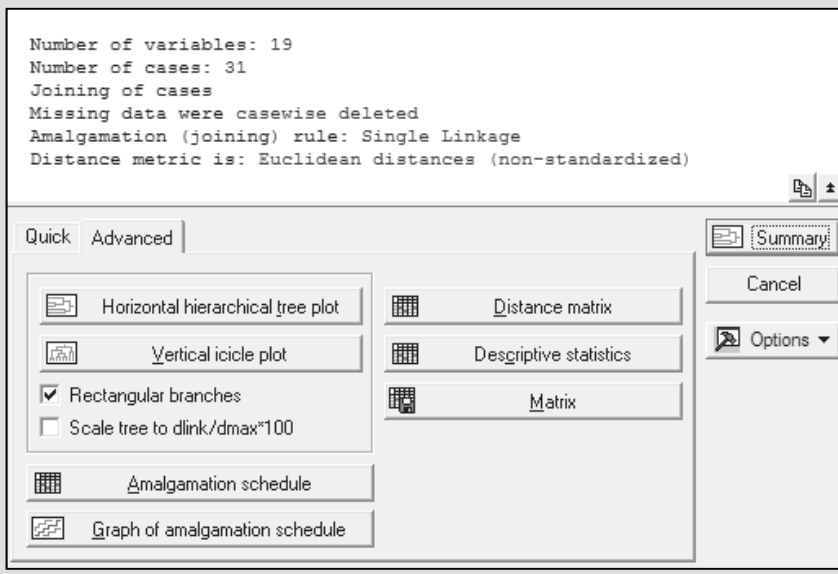


Paso 4. En la ventana anterior, al pulsar en «Variables», aparece el siguiente cuadro de diálogo donde se seleccionan todos los pigmentos.

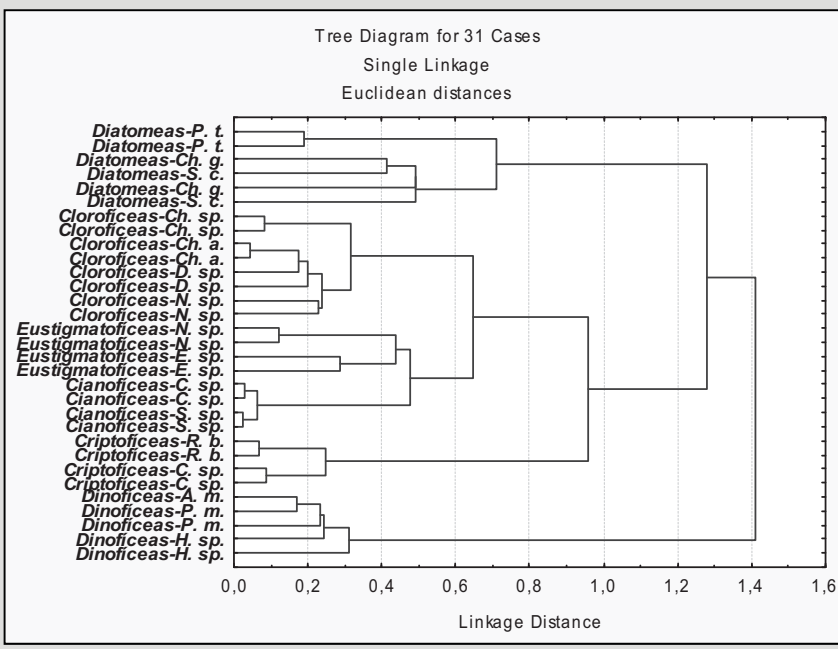


CUADRO XIV.1. *(Continuación)*[STATISTICA]

Paso 5. Por último, pulsamos en «Horizontal hierarchical tree plot» y aparece el dendrograma.



Paso 6. Se observa que las diferentes clases de algas se agrupan perfectamente en función del tipo de pigmentos que tienen. Las Diatomeas forman un grupo muy diferente del resto, así como Criptofíceas y Dinofíceas. Las clases con mayor parecido en los pigmentos son las Eustigmatofíceas y Cianofíceas.

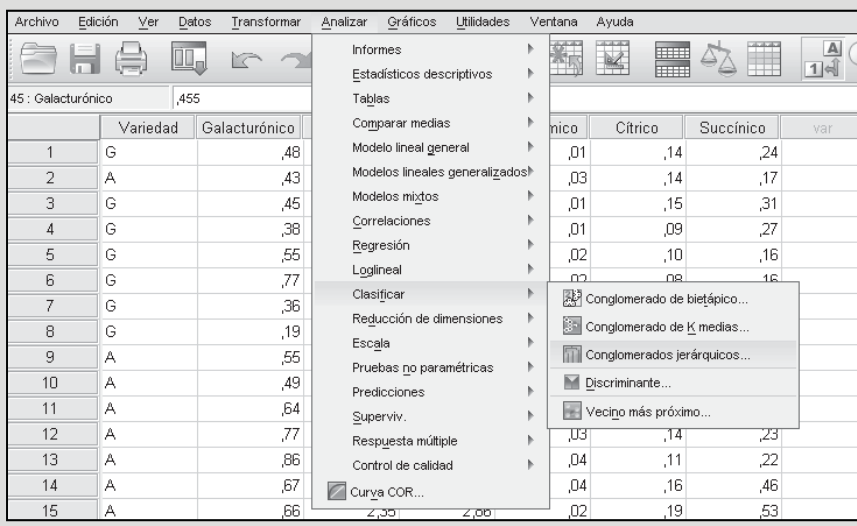


CUADRO XIV.1. *(Continuación)*[SPSS]

Aplicación con SPSS.

EJEMPLO. Utilizaremos los mismos datos del ejemplo con R, relativo a una muestra de vinos elaborados con dos variedades distintas de uva en los que se han determinado las concentraciones de 6 ácidos orgánicos. Los datos están en el archivo **Cuadro XIV.1.sav**.

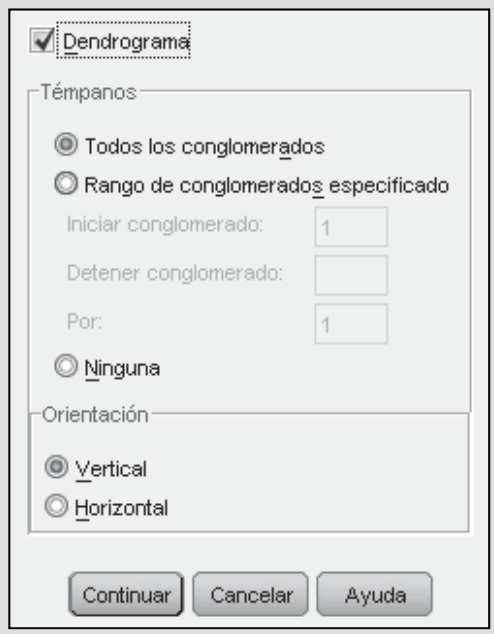
Paso 1. Se selecciona «Analizar» ⇨ «Clasificar» ⇨ «Conglomerados jeráquicos…».



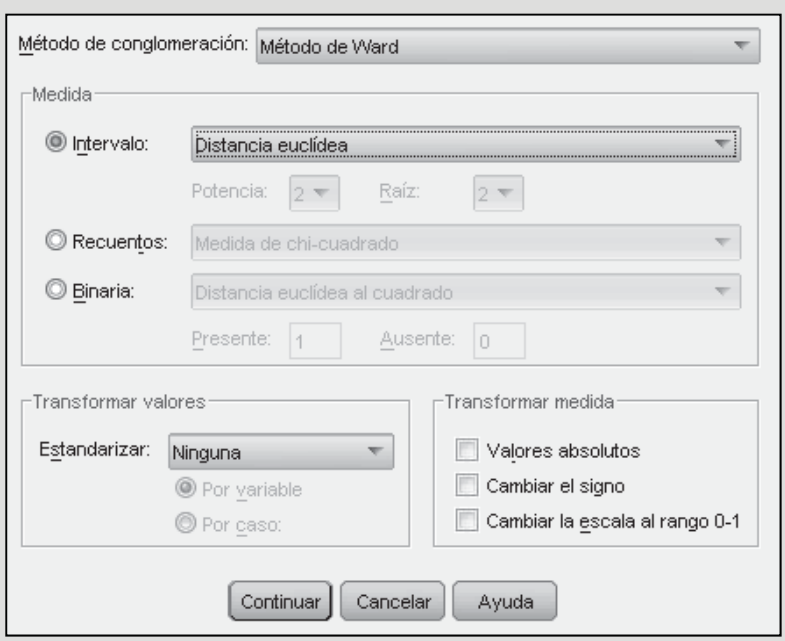
Paso 2. En el cuadro de diálogo que aparece seleccionamos las variables que queremos utilizar, en este caso las seis concentraciones, y la variable «Variedad» como etiqueta de caso, para identificar la variedad de uva en el dendrograma. La variable que se usa como etiqueta debe ser de cadena y no numérica.

CUADRO XIV.1. *(Continuación)*[SPSS]

Paso 3. En la ventana anterior pulsamos «Gráficos» y en el cuadro que aparece marcamos «Dendrograma».



Paso 4. En la ventana del paso 2 pulsamos «Método» y aparece el cuadro de diálogo donde podemos especificar el método de conglomeración y la distancia.



CUADRO XIV.1. *(Continuación)*[SPSS]

El «Método de Ward» puede ser adecuado cuando buscamos clases con poca variabilidad interna (con elementos muy similares entre sí). Elegimos también la distancia apropiada para nuestros datos: como se trata de variables cuantitativas empleamos el primer desplegable (escala de intervalo) y seleccionamos «Distancia euclídea».

Otros desplegables permiten elegir distancias más adecuadas para datos que son recuentos o frecuencias, e incluso para datos de respuesta binaria.

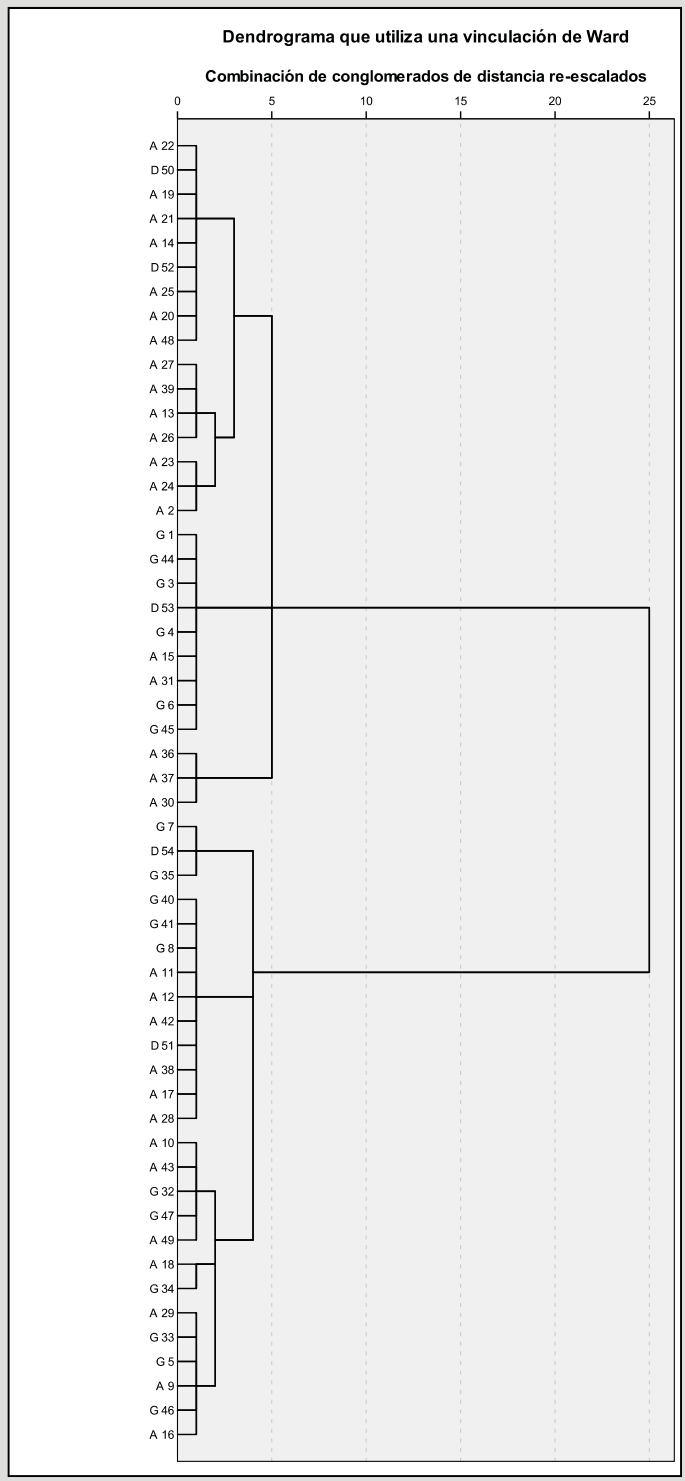
También es posible transformar los valores (estandarizar, adecuado para datos heterogéneos que vienen dados en diferentes unidades de medida) o transformar las distancias: por ejemplo, se puede utilizar el coeficiente de correlación como medida de distancia entre variables, en cuyo caso puede ser útil emplear el valor absoluto, ya que tanto los coeficientes grandes positivos como los negativos indican variables similares, y también puede ser conveniente cambiar el signo, ya que los valores altos no indican más distancia, sino más proximidad entre variables, y los pequeños más diferencia.

Paso 5. En la siguiente página se muestra el gráfico. Las etiquetas de caso «A» (Albariño) y «G» (Godello) permiten identificar la variedad de uva utilizada en cada elemento de la muestra, y relacionar esta característica con la clasificación obtenida. En las dos grandes clases, en el primer nivel de clasificación, ambas variedades aparecen entremezcladas, lo que parece indicar que la variedad de uva no es el principal factor de diferencia entre los elementos de nuestra muestra, sino quizás la zona de procedencia, u otra característica no registrada.

Los elementos con variedad «D» (desconocida) se sitúan en diferentes grupos. En algunos solamente hay elementos de la misma variedad, como los dos primeros en la parte superior del gráfico, situados en una clase en la que solamente hay Albariño, por lo que podemos utilizar el dendrograma para identificar –por afinidad– la variedad de uva desconocida. En otros sin embargo hay mezcla de variedades y no es posible la identificación por esta vía.

La clasificación principal en el dendrograma es aquella que muestra la mayor distancia posible entre las clases, medida en la escala horizontal en el gráfico anterior. En nuestro ejemplo es la que tiene dos clases; como podemos comprobar, las distancias entre elementos dentro de cada una de estas dos clases son mucho menores que las distancias entre las clases. De este modo las clases son homogéneas internamente, pero muy diferentes entre sí.

CUADRO XIV.1. *(Continuación)*[SPSS]



XIV.2. Clasificación de *k*-medias

Este método trata de encontrar la mejor partición posible del conjunto de objetos para un número dado de clases *k*.

El algoritmo comienza con el establecimiento de los centros iniciales de clase, que pueden ser algunos elementos de la muestra convenientemente separados entre sí, o bien elementos elegidos al azar. A continuación se asignan todos los elementos a la clase cuyo centro está más próximo y se recalculan los centros como la media de los elementos de cada clase, iterando estos dos pasos -asignación y cálculo de centros- hasta que la clasificación obtenida es estable. La regla de parada del algoritmo es múltiple: se termina cuando se repite la clasificación, cuando se ha alcanzado el número máximo de iteraciones establecido a priori, o bien cuando el cambio en el desplazamiento de los centros de clase entre dos iteraciones consecutivas es muy pequeño, menor que un valor umbral predeterminado.

El algoritmo reduce la varianza interna de las clases en cada iteración, haciendo por lo tanto que las clases sean cada vez más homogéneas. Esta es una característica importante de la clasificación que se busca.

Con este método –a diferencia de la clasificación jerárquica– es necesario establecer a priori el número de clases que debe tener la solución buscada. Sin embargo se pueden obtener distintas clasificaciones, con distinto número de grupos, aplicando el método de forma repetida.

El resultado puede depender en algunos casos de los centros iniciales elegidos, aunque en general la solución final es muy robusta frente a esta elección, y suele ser prácticamente independiente de los centros utilizados para iniciar el algoritmo. Algunos programas estadísticos utilizan diferentes arranques aleatorios y comprueban después si la solución obtenida es la misma con todos ellos.

Dado que el procedimiento consiste en reducir paulatinamente la variabilidad interna, el método obtiene clases con dispersión aproximadamente esférica. No es adecuado por lo tanto en aquellas aplicaciones en las que sabemos que las clases no tienen esa forma.

Es adecuado para clasificar conjuntos de elementos muy numerosos (por ejemplo con cientos de miles de objetos), con los cuales el método de clasificación jerárquica es inviable dadas las dificultades de cálculo.

Se suele utilizar a continuación del método jerárquico, sirviendo éste para decidir el número de clases que se desea obtener; si el conjunto de objetos es muy grande, es suficiente para este propósito –determinación del número de clases– una muestra aleatoria.

Entre los resultados de la aplicación del método de clasificación de k-medias, en general se obtienen los siguientes:

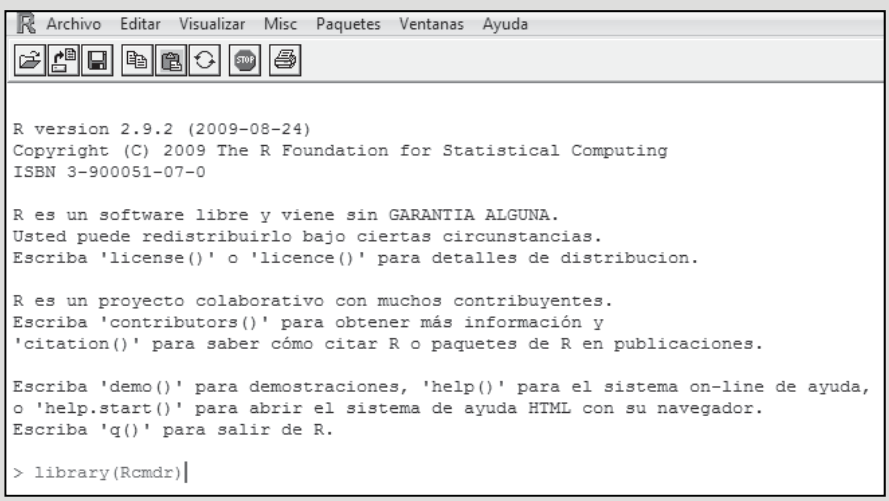
- *Número de elementos en cada clase*.
- *Centros de clase*: permiten identificar las características principales de las clases, interpretando los valores medios.
- *Varianza interna y total*: cuanto más pequeña sea la primera en relación a la segunda, mejor es la clasificación. También se suele mostrar la reducción de la varianza interna en cada paso o iteración del algoritmo.

CUADRO XIV.2. *Clasificación de k-medias*

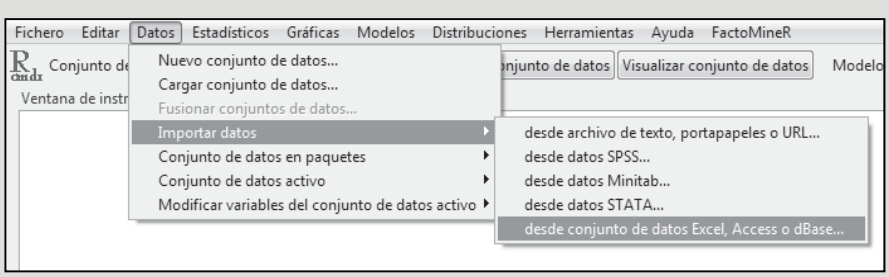
Aplicación con R.

EJEMPLO. Utilizaremos el mismo ejemplo que en el Cuadro XIV.1, lo cual nos permitirá ver mejor las diferencias entre ambos métodos de clasificación. Los datos, que están en el archivo **Cuadro XIV.1.xls**, son concentraciones de diferentes ácidos orgánicos (galacturónico, tartárico, málico, shiquímico, cítrico, succínico) en dos tipos de vino elaborados con distintas uvas: Albariño (A) y Godello (G). El objetivo es averiguar si nuestros vinos se agrupan en clases razonablemente homogéneas, que pudieran corresponderse con la variedad de uva utilizada.

Paso 1. Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, cargamos el paquete "Rcmdr" con la instrucción library(Rcmdr). Si el paquete no está instalado, la instalación se realiza como se explica en el Apéndice II.

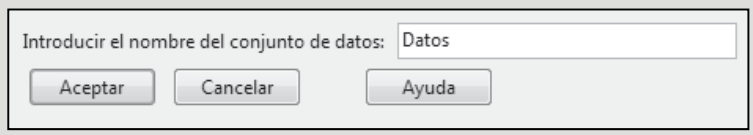

Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> library(Rcmdr)

Paso 2. Se selecciona «Datos» ⇨ «Importar datos» ⇨ «desde conjunto de datos Excel, Access o dBase».


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda FactoMineR
Nuevo conjunto de datos...
Cargar conjunto de datos...
Fusionar conjuntos de datos...
Importar datos
Conjunto de datos en paquetes
Conjunto de datos activo
Modificar variables del conjunto de datos activo
desde archivo de texto, portapapeles o URL...
desde datos SPSS...
desde datos Minitab...
desde datos STATA...
desde conjunto de datos Excel, Access o dBase...
Visualizar conjunto de datos
Modelo

CUADRO XIV.2. *(Continuación)*[R]

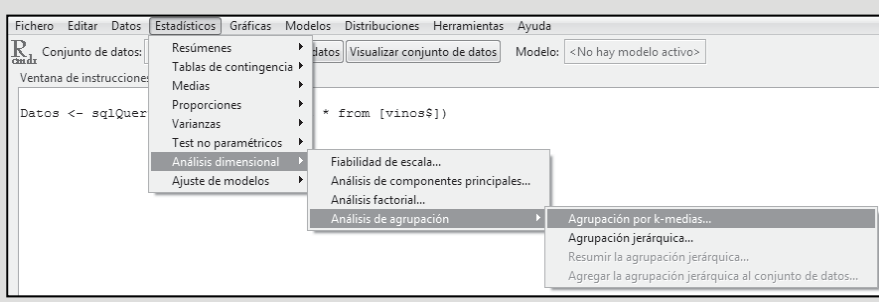
Paso 3. Aparece un pequeño cuadro de diálogo para introducir el nombre del conjunto de datos, en el que dejamos el nombre por defecto «Datos» y pulsamos «Aceptar». Posteriormente aparecen los archivos del directorio de trabajo, en el que elegimos **Cuadro XIV.1.xls** y pulsamos «Abrir».



Paso 4. En R Commander pulsamos en «Visualizar conjunto de datos». Es importante inspeccionar los datos y que no existan líneas, que a veces ocurre al final del archivo, que tengan celdas con el código <NA>. Si existen estas líneas, hay que eliminarlas.

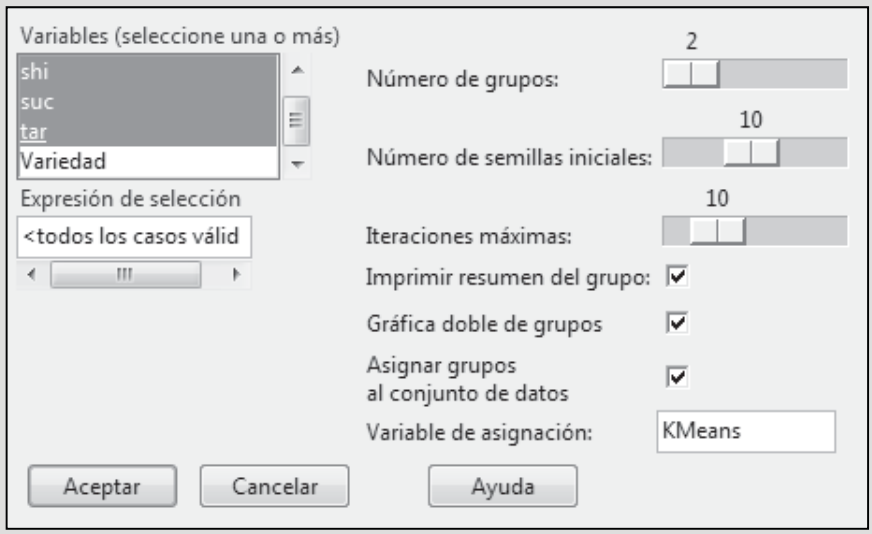
	Variedad	Etiqueta	gal	tar	mal	shi	cit	suc
1	2	G	0.482	1.796	2.562	0.006	0.136	0.236
2	1	A	0.429	2.749	3.458	0.032	0.138	0.170
3	2	G	0.453	2.193	2.716	0.007	0.152	0.311
4	2	G	0.375	2.048	3.055	0.007	0.094	0.268
5	2	G	0.548	1.977	1.540	0.018	0.096	0.156
6	2	G	0.768	2.462	2.625	0.019	0.080	0.158
7	2	G	0.356	0.726	0.478	0.019	0.046	0.329

Paso 5. Para aplicar el método de clasificación k-medias elegimos en el menú principal de R Commander las opciones «Estadísticos» ⇨ «Análisis dimensional» ⇨ «Análisis de agrupación» ⇨ «Agrupación por k-medias…».



Paso 6. En la ventana de agrupación por k-medias seleccionamos en primer lugar las variables que queremos utilizar: todas excepto «Variedad». Estableceremos el «Número de grupos:» (2), ya que estamos trabajando con 2 tipos de vino: Albariño y Godello. En general pondremos el número de grupos que esperamos encontrar. Se mantienen los valores por defecto en el resto de los parámetros: «Número de semillas iniciales» (10) e «Iteraciones máximas» (10). Dejaremos la marca en «Imprimir resumen del grupo:» y «Gráfica doble de grupos», y marcamos «Asignar grupos al conjunto de datos».

CUADRO XIV.2. *(Continuación)*[R]



Las semillas iniciales se utilizan para determinar los centros iniciales aleatoriamente. Cada una de las 10 utilizadas da origen a una clasificación diferente, lo que permite valorar la influencia que tiene esta elección de centros iniciales sobre el resultado final. En general el efecto es nulo, y la clasificación obtenida es la misma.

El número máximo de 10 iteraciones generalmente no se alcanza, ya que antes de que eso ocurra suele repetirse la clasificación, con lo que finaliza el algoritmo. Si no se alcanza la convergencia, generalmente porque no existe una clasificación natural bien definida en la muestra, puede repetirse el método aumentando el número de iteraciones (por ejemplo 20).

Paso 7. Interpretamos los resultados que se obtienen. Es importante mencionar que, como hemos decidido hacer sólo 2 grupos, el análisis establece arbitrariamente los grupos 1 y 2, los cuales no se corresponden con las variedades de vinos 1 y 2. Dentro del grupo 1 pueden estar las dos variedades y lo mismo en el grupo 2.

a) Tamaño –número de elementos– de cada grupo: 28 el grupo 1 y 26 el grupo 2. Como asigna arbitrariamente los grupos, es posible que intercambie las etiquetas y salga al revés, es decir 26 grupo 1 y 28 grupo 2. Lo mismo ocurre con el resto de resultados.

b) Centroides de grupo, o valores medios de las variables en cada grupo. Se puede comprobar como en el grupo 1 las concentraciones de todos los ácidos orgánicos son mayores, y la diferencia entre grupos es más alta en el ácido málico.

	cit	gal	mal	shi	suc	tar
1	0.171	0.623	3.346	0.027	0.382	2.003
2	0.130	0.529	0.950	0.024	0.316	1.863

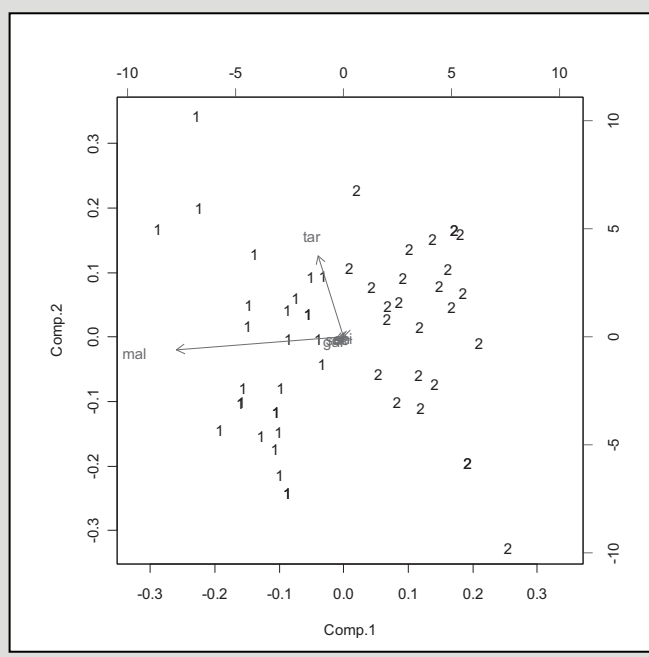
CUADRO XIV.2. *(Continuación)*[R]

c) Suma de cuadrados dentro de cada grupo: 23.68309 en el 1 y 20.95674 en el grupo 2

d) Suma de cuadrados externa o entre grupos: 77.85233

Estos dos últimos resultados, las sumas interna y externa se utilizan para evaluar la clasificación resultante: cuanto más pequeña sea la suma interna comparada con la externa, mejor será el resultado, ya que la suma interna mide la dispersión o distancia entre los elementos del mismo grupo, y la externa mide la distancia entre grupos. Lo óptimo es un conjunto de grupos muy homogéneos y muy separados entre sí.

Paso 8. El gráfico representa los elementos de los dos grupos –identificados con el número de grupo (1,2)– conjuntamente con las variables (en color rojo). Ya se mencionó anteriormente que la asignación arbitraria puede generar que las etiquetas de los grupos salgan invertidas. Solamente dos variables tienen una influencia importante en la clasificación, tartárico y málico, y la segunda caracteriza al grupo 1. El resto de variables, situadas todas ellas en el centro del gráfico, tienen un efecto pequeño sobre la formación de los grupos. Vemos por otra parte que los grupos están próximos entre sí, y tienen una notable dispersión interna, de modo que algunos elementos están más próximos a elementos de la otra clase que a muchos de su propio grupo.



Paso 9. Si queremos saber cuales son los elementos de la muestra asignados a cada una de las dos clases, solo tenemos que inspeccionar de nuevo nuestros datos pulsando en «Visualizar conjunto de datos». Se ha añadido una variable nueva, en último lugar a la derecha, denominada «Kmeans», con el número de grupo para cada elemento de la muestra.

CUADRO XIV.2. *(Continuación)*[R]

Por ejemplo el primer elemento, que es un vino de Godello, el análisis lo asigna al grupo 1 y el segundo elemento, que es de Albariño, también lo asigna al grupo 1.

	Variedad	Etiqueta	gal	tar	mal	shi	cit	suc	KMeans
1	2	G	0.482	1.796	2.562	0.006	0.136	0.236	1
2	1	A	0.429	2.749	3.458	0.032	0.138	0.170	1
3	2	G	0.453	2.193	2.716	0.007	0.152	0.311	1
4	2	G	0.375	2.048	3.055	0.007	0.094	0.268	1
5	2	G	0.548	1.977	1.540	0.018	0.096	0.156	2

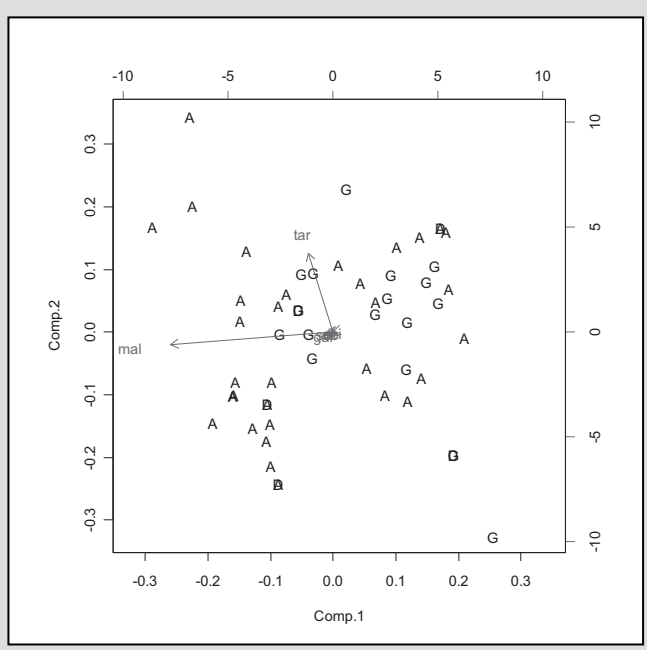
Paso 10. Es posible añadir al gráfico biplot etiquetas de caso, con la variedad de uva utilizada (o cualquier otra característica disponible en los datos). Para ello solo tenemos que modificar en la ventana de instrucciones de R Commander la orden:

biplot(princomp(model.matrix(~-1 + cit + gal + mal + shi + suc + tar, Datos)), xlabs = **as.character(.cluster$cluster)**)

por

biplot(princomp(model.matrix(~-1 + cit + gal + mal + shi + suc + tar, Datos)), xlabs = **Datos[,2]**)

que ejecutaremos seleccionando la instrucción corregida completa y pulsando después «Ejecutar». Ahora podemos observar como, aunque existe una mezcla de variedades de uva en ambas clases, las "G" predominan claramente en algunas zonas del gráfico, y en otras las "A", y en consecuencia la variedad de uva (Godello, Albariño) tiene una cierta influencia en la agrupación.

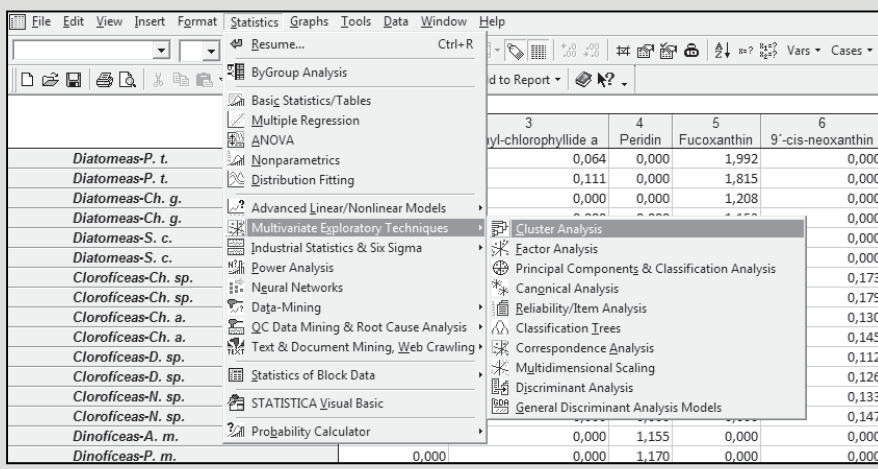


CUADRO XIV.2. *(Continuación)*[STATISTICA]

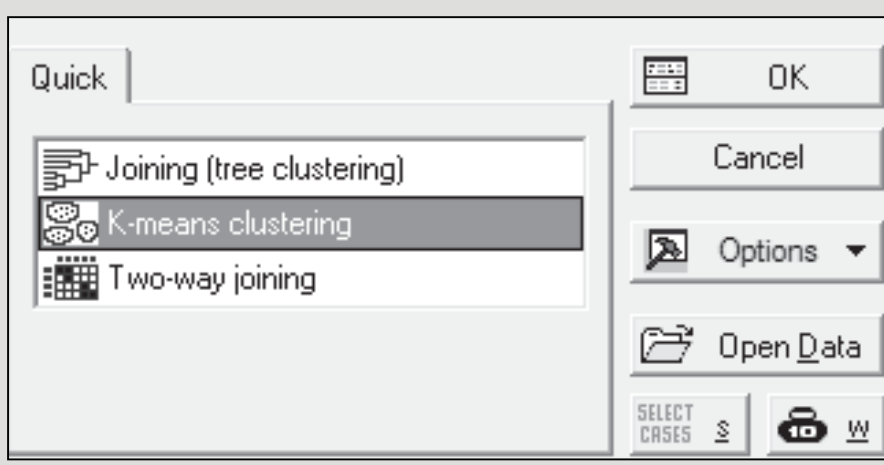
Aplicación con STATISTICA.

EJEMPLO. Para ver mejor la diferencia entre el método de clasificación jerárquica y el de k-medias usaremos los datos del Cuadro XIV.1 que están en el archivo **Cuadro XIV.1.sta**. El objetivo es determinar si es posible diferenciar las clases de algas (Diatomeas, Clorofíceas, Dinofíceas, Eustigmatofíceas, Cianofíceas y Criptofíceas) en función de su composición relativa de pigmentos.

Paso 1. Se selecciona «Statistics» ⇨ «Multivariate Exploratory Techniques» ⇨ «Cluster Analysis».

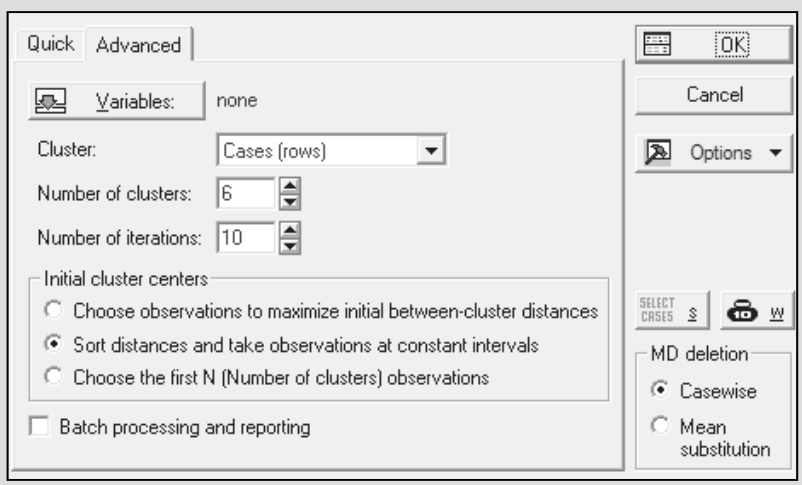


Paso 2. A continuación «K-means clustering».

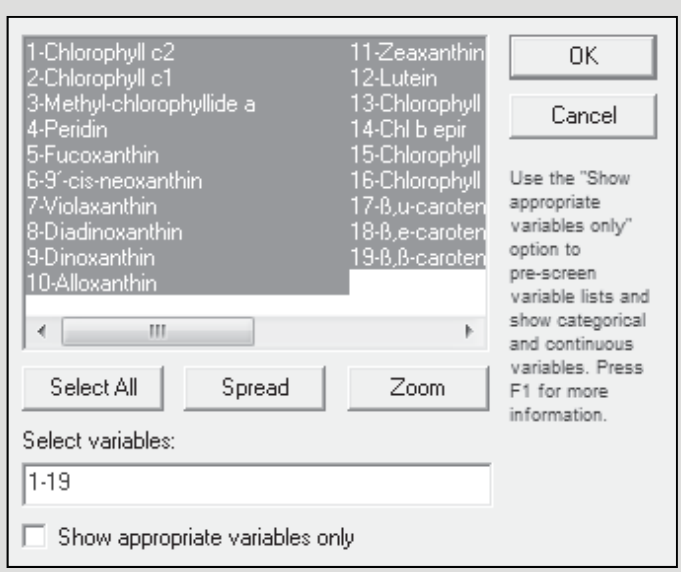


CUADRO XIV.2. *(Continuación)*[STATISTICA]

Paso 3. En la ventana que aparece pulsamos en la pestaña «Advanced». En «Cluster» elegiremos clasificar los casos «Cases (rows)» en lugar de las variables, es decir, realizar el análisis para ver la asociación entre las clases de algas. Como hay 6 clases de algas, ese será el valor que pondremos en «Number clusters:». Conservamos las opciones por defecto «Number of iterations:» igual a 10 y marcamos «Sort distances and take observations at constant intervals». También mantenemos la opción «Casewise» para el tratamiento de los valores perdidos (excluirá todos los casos en los que falte algún dato) en lugar de «Mean substitution». Esta última opción, pese a la distorsión que supone, puede ser adecuada cuando existen muchos valores faltantes en nuestros datos.

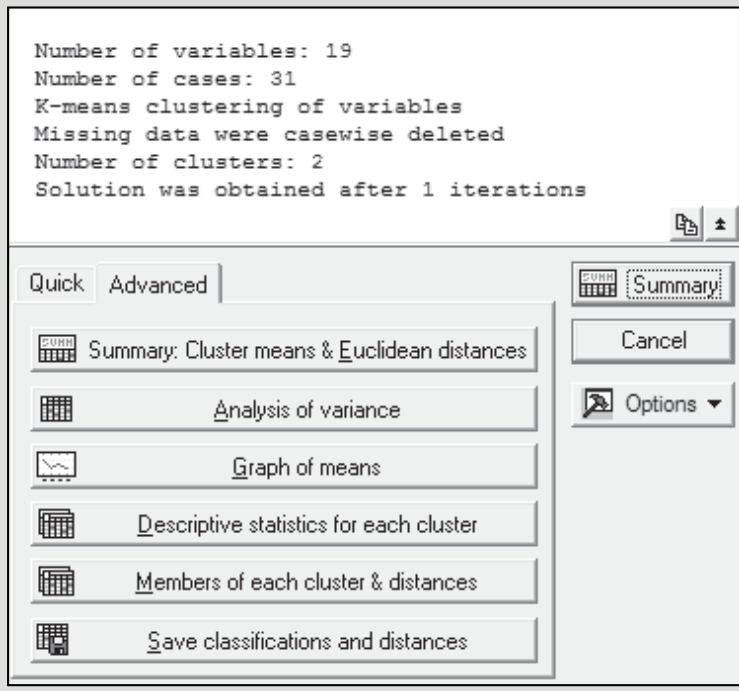


Paso 4. En la ventana anterior, al pulsar en «Variables», aparece el siguiente cuadro de diálogo donde se seleccionan todos los pigmentos.



CUADRO XIV.2. *(Continuación)*[STATISTICA]

Paso 5. En la pestaña «Advanced» pulsamos en «Summary: Cluster means & Euclidean distances» y obtenemos las medias de cada grupo.



	Cluster Means (Cuadro XIV.1.sta)					
Variable	Cluster No. 1	Cluster No. 2	Cluster No. 3	Cluster No. 4	Cluster No. 5	Cluster No. 6
Chlorophyll c2	0,886464	0,000000	0,000000	0,418712	0,499509	0,488062
Chlorophyll c1	0,000000	0,000000	0,000000	0,111437	0,000000	0,510889
Methyl-chlorophyllide a	0,000000	0,012213	0,003461	0,087372	0,000000	0,020848
Peridin	1,195013	0,000000	0,000000	0,000000	0,000000	0,000000
Fucoxanthin	0,000000	0,000000	0,000000	1,903305	0,000000	1,201748
9´-cis-neoxanthin	0,000000	0,000000	0,143175	0,000000	0,000000	0,000000
Violaxanthin	0,000000	0,320238	0,061886	0,000000	0,000000	0,000000
Diadinoxanthin	0,136755	0,000000	0,000000	0,000000	0,000000	0,000000
Dinoxanthin	0,000000	0,080780	0,008830	0,015386	0,000000	0,003295
Alloxanthin	0,000000	0,000000	0,000000	0,000000	0,943129	0,000000
Zeaxanthin	0,000000	0,227399	0,063794	0,000000	0,000000	0,000000
Lutein	0,000000	0,001816	0,732202	0,000000	0,000000	0,000000
Chlorophyll b	0,000000	0,000000	0,226253	0,000000	0,000000	0,000000
Chl b epir	0,000000	0,000000	0,003662	0,000000	0,000000	0,000000
Chlorophyll a alllomer	0,019787	0,056064	0,038839	0,000000	0,046423	0,088609
Chlorophyll a epimer	0,027978	0,030144	0,014936	0,028673	0,023711	0,017097
ß,u-carotene	0,000000	0,000000	0,000000	0,000000	0,002562	0,000000
ß,e-carotene	0,000000	0,000000	0,042861	0,000000	0,109698	0,000000
ß,ß-carotene	0,067034	0,176093	0,066694	0,096334	0,015246	0,017629

El grupo 1 se caracteriza por los valores altos del pigmento Peridina, el grupo 2 por Violaxantina, el 3 por la Luteina, el 4 por la Fucoxantina, el 5 por la Aloxantina y, el 6 por la Fucoxantina. Por lo tanto, cada grupo parece tener claramente asociado un tipo de pigmento.

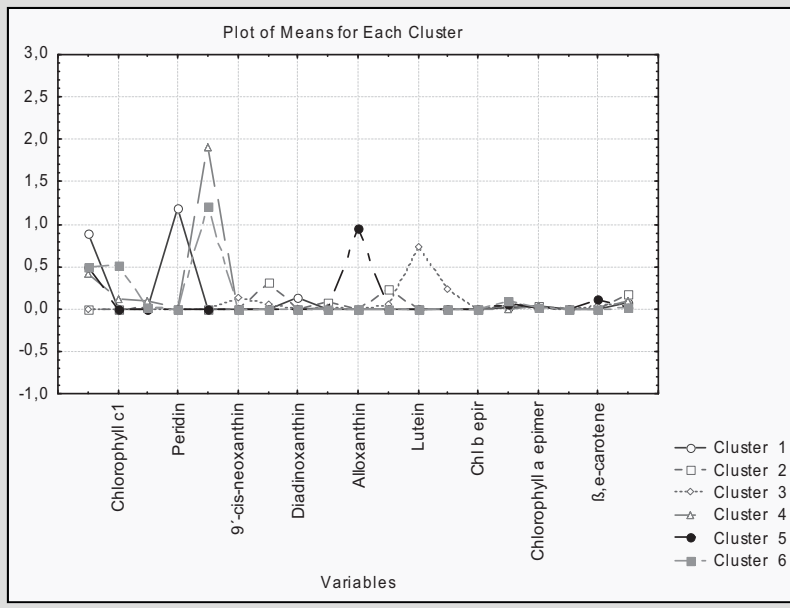
CUADRO XIV.2. *(Continuación)*[STATISTICA]

Paso 6. Si en la ventana del paso 5 pulsamos en «Analysis of variance» obtenemos una tabla que presenta, para cada variable, la suma de cuadrados de desviaciones dentro de las clases «Within» y entre clases «Between», y el cociente entre ambos cuadrados medios «F», tras dividir por los grados de libertad. Los valores grandes de «F» indican las variables que toman valores distintos en los grupos y, por lo tanto, las que caracterizan mejor a las clases y explican sus diferencias. No debe ser interpretado el valor p, que también se muestra, ya que en general no se cumplen las condiciones para poder utilizarlo. Es suficiente para que «F» sea alto con que alguna clase tenga una media muy distinta de otras. Por ejemplo, 9´-cis-neoxantina y Clorofila b en la tabla de medias sus valores no son muy grandes, pero tienen valores de «F» altos, porque ambas variables destacan en el grupo 3.

	Analysis of Variance (Cuadro XIV.1.sta)					
Variable	Between SS	df	Within SS	df	F	signif. p
Chlorophyll c2	3,48836	5	0,399760	25	43,6307	0,000000
Chlorophyll c1	0,90317	5	0,170065	25	26,5537	0,000000
Methyl-chlorophyllide a	0,01355	5	0,008456	25	8,0121	0,000127
Peridin	5,98863	5	0,080126	25	373,6992	0,000000
Fucoxanthin	10,62858	5	0,092819	25	572,5443	0,000000
9´-cis-neoxanthin	0,12167	5	0,003727	25	163,2441	0,000000
Violaxanthin	0,54960	5	1,201803	25	2,2866	0,076676
Diadinoxanthin	0,07843	5	0,017244	25	22,7400	0,000000
Dinoxanthin	0,03467	5	0,077240	25	2,2444	0,081213
Alloxanthin	3,09888	5	0,047737	25	324,5783	0,000000
Zeaxanthin	0,27118	5	0,232239	25	5,8384	0,001050
Lutein	3,17666	5	0,271008	25	58,6082	0,000000
Chlorophyll b	0,30384	5	0,018267	25	83,1676	0,000000
Chl b epir	0,00008	5	0,000086	25	4,6185	0,004027
Chlorophyll a alllomer	0,01613	5	0,045597	25	1,7684	0,156054
Chlorophyll a epimer	0,00126	5	0,005921	25	1,0606	0,405382
ß,u-carotene	0,00002	5	0,000034	25	3,3231	0,019467
ß,e-carotene	0,04312	5	0,065138	25	3,3099	0,019798
ß,ß-carotene	0,10852	5	0,168448	25	3,2211	0,022188

Paso 7. Al pulsar «Graph of means» en la ventana del paso 5, se obtiene el gráfico donde se observa la contribución de cada pigmento a cada grupo, es decir, las medias que se mostraron en la tabla del paso 5.

CUADRO XIV.2. *(Continuación)*[STATISTICA]



Paso 8. Al pulsar «Members of cluster & distances» en la ventana del paso 5, podemos ver cuáles son las clases de algas que integran cada grupo. Por ejemplo, el grupo 1 son las Dianofíceas. En el grupo 2 hay dos clases: Cianofíceas y Eustigmatofíceas. En el grupo 3 solo hay Clorofíceas. En el grupo 4 una especie de Diatomeas. En el grupo 5 sólo Criptofíceas. Por último, en el grupo 6 el resto de especies de Diatomeas. Por tanto, los pigmentos identifican bastante bien la clase de alga, porque la mayoría de los grupos están representados por un solo tipo de alga.

	Members of Cluster Number 1 (Cuadro XIV.1.sta) and Distances from Respective Cluster Center Cluster contains 5 cases
	Distance
Dinoficeas-A. m.	0,045668
Dinoficeas-P. m.	0,028132
Dinoficeas-P. m.	0,047587
Dinoficeas-H. sp.	0,045922
Dinoficeas-H. sp.	0,057937

	Members of Cluster Number 2 (Cuadro XIV.1.sta) and Distances from Respective Cluster Center Cluster contains 8 cases
	Distance
Eustigmatoficeas-N. sp.	0,046774
Eustigmatoficeas-N. sp.	0,059869
Eustigmatoficeas-E. sp.	0,180204
Eustigmatoficeas-E. sp.	0,119161
Cianoficeas-C. sp.	0,091575
Cianoficeas-C. sp.	0,089631
Cianoficeas-S. sp.	0,083479
Cianoficeas-S. sp.	0,085470

	Members of Cluster Number 3 (Cuadro XIV.1.sta) and Distances from Respective Cluster Center Cluster contains 8 cases
	Distance
Cloroficeas-Ch. sp.	0,071580
Cloroficeas-Ch. sp.	0,087869
Cloroficeas-Ch. a.	0,050796
Cloroficeas-Ch. a.	0,045051
Cloroficeas-D. sp.	0,054644
Cloroficeas-D. sp.	0,044895
Cloroficeas-N. sp.	0,056898
Cloroficeas-N. sp.	0,023153

	Members of Cluster Number 4 (Cuadro XIV.1.sta) and Distances from Respective Cluster Center Cluster contains 2 cases
	Distance
Diatomeas-P. t.	0,021677
Diatomeas-P. t.	0,021677

	Members of Cluster Number 5 (Cuadro XIV.1.sta) and Distances from Respective Cluster Center Cluster contains 4 cases
	Distance
Criptoficeas-R. b.	0,041315
Criptoficeas-R. b.	0,031434
Criptoficeas-C. sp.	0,026181
Criptoficeas-C. sp.	0,045621

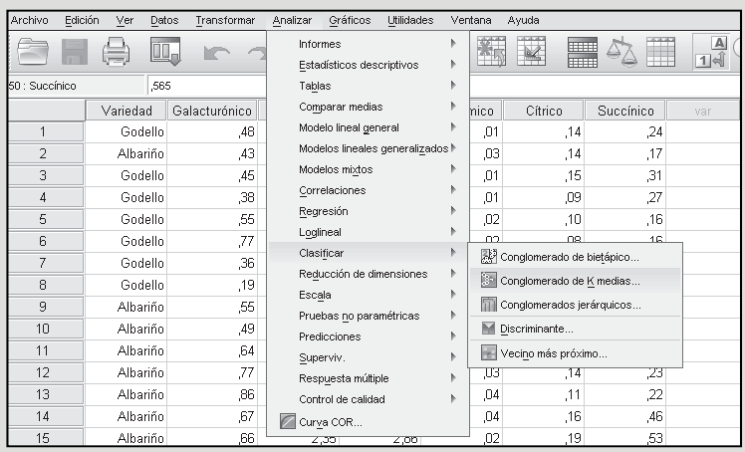
	Members of Cluster Number 6 (Cuadro XIV.1.sta) and Distances from Respective Cluster Center Cluster contains 4 cases
	Distance
Diatomeas-Ch. g.	0,066806
Diatomeas-Ch. g.	0,093796
Diatomeas-S. c.	0,064173
Diatomeas-S. c.	0,095796

CUADRO XIV.2. *(Continuación)*[SPSS]

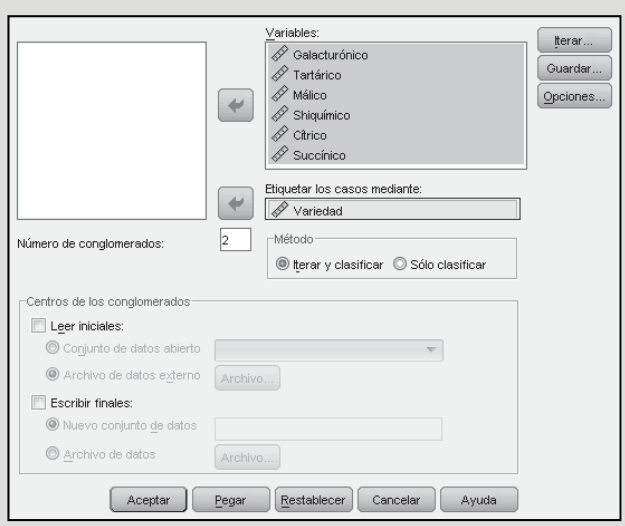
Aplicación con SPSS.

EJEMPLO. Utilizaremos los mismos datos del Cuadro XIV.1 en la aplicación con R, relativos a una muestra de vinos elaborados con dos variedades de uvas distintas en los que se han determinado las concentraciones de 6 ácidos orgánicos. Los datos están en el archivo **Cuadro XIV.1.sav**.

Paso 1. Se selecciona «Analizar» ⇨ «Clasificar» ⇨ «Conglomerados de K medias…».

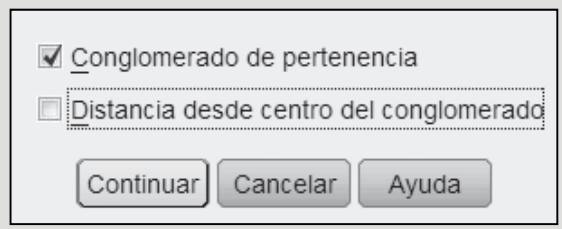


Paso 2. En el cuadro de diálogo que aparece seleccionamos las variables que queremos utilizar, en este caso las seis concentraciones, y la variable «Variedad» como etiqueta de caso, para identificar la variedad de uva en los resultados. La variable que se usa como etiqueta debe ser de cadena y no numérica. Se indica 2 en «Número de conglomerados», ya que sólo hay 2 tipos de vinos (Albariño y Godello).

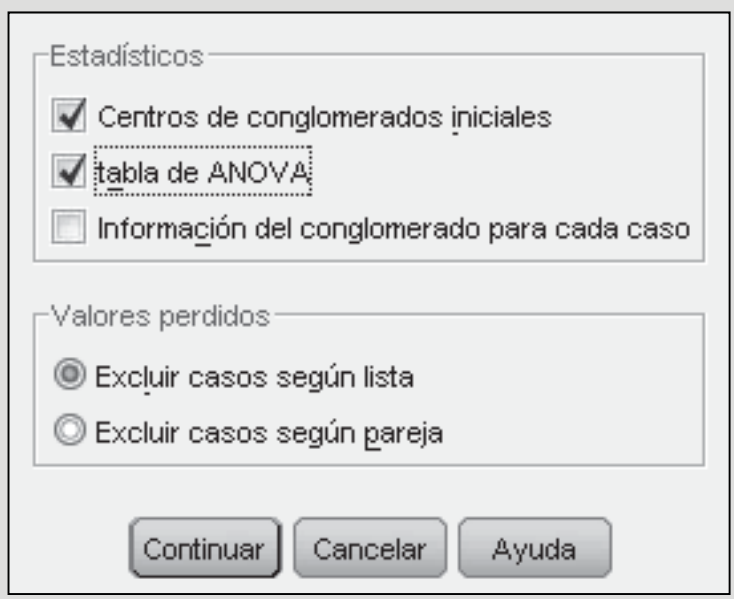


CUADRO XIV.2. *(Continuación)*[SPSS]

Paso 3. En el cuadro de diálogo «Guardar» marcamos «Conglomerado de pertenencia».



Paso 4. En el cuadro de diálogo «Opciones» marcamos «Centros de conglomerados iniciales» y «tabla de ANOVA».



Paso 5. En los resultados, en primer lugar analizamos el historial de iteraciones, que nos indica que han sido necesarias 4 iteraciones, y el algoritmo se detiene por repetición de la clasificación, ya que el cambio en los centros en la iteración 4 es nulo.

Iteración	Cambio en los centros de los conglomerados	
	1	2
1	2,088	1,991
2	,062	,059
3	,053	,051
4	,000	,000

CUADRO XIV.2. *(Continuación)*[SPSS]

Paso 6. Los centros finales de grupo o centroides son los valores medios de las variables en cada grupo. Los dos grupos se diferencian principalmente por el ácido málico.

Centros de los conglomerados finales

	Conglomerado	
	1	2
gal	,529	,624
tar	1,863	2,003
mal	,951	3,347
shi	,024	,027
cit	,131	,171
suc	,316	,382

Paso 7. Los resultados del ANOVA muestran el cuadrado medio entre grupos «Conglomerado» y dentro de los grupos «Error», y el cociente entre ambos «*F*». En general no debe ser interpretado el valor *p*, que también se muestra, ya que no suelen cumplirse las condiciones o hipótesis para ello. Los valores grandes de *F* indican variabilidad entre grupos mayor que dentro de los grupos, es decir grupos homogéneos y distintos entre sí, con respecto a esa variable. Son por lo tanto las variables que marcan la diferencia entre las clases.

ANOVA

	Conglomerado		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
gal	,121	1	,036	52	3,387	,071
tar	,264	1	,466	52	,567	,455
mal	77,387	1	,329	52	235,545	,000
shi	,000	1	,000	52	,796	,376
cit	,022	1	,005	52	4,137	,047
suc	,059	1	,023	52	2,562	,116

Paso 8. En la matriz de datos se ha generado una nueva columna «QCL_1», que indica el grupo de pertenencia de cada caso. Por ejemplo, el primer vino que era de Godello se asignó al grupo 2, aunque el segundo vino, el cual es de Albariño, también fue asignado al grupo 2. El 64,1% de los vinos Godello están en el grupo 1 y el 59,4% de los vinos de Albariño están en el grupo 2. Por tanto, los ácidos orgánicos de cada tipo de variedad de uva (Godello, Albariño) no permiten diferenciar de forma clara entre ambos tipos de vino.

	Variedad	Galacturónico	Tartárico	Málico	Shiquímico	Cítrico	Succínico	QCL_1
1	Godello	,48	1,80	2,56	,01	,14	,24	2
2	Albariño	,43	2,75	3,46	,03	,14	,17	2
3	Godello	,45	2,19	2,72	,01	,15	,31	2
4	Godello	,38	2,05	3,06	,01	,09	,27	2
5	Godello	,55	1,98	1,54	,02	,10	,16	1
6	Godello	,77	2,46	2,63	,02	,08	,16	2

XIV.3. Análisis Discriminante

Una técnica estadística multivariante ampliamente utilizada para caracterizar clases de elementos cuando se dispone de un cierto número de variables es el Análisis Discriminante, que permite -a partir de una clasificación inicial- obtener una nueva clasificación en función de los valores observados de un conjunto de variables. El método consiste en la determinación de unas funciones de las variables originales, llamadas funciones discriminantes, que permiten decidir en qué clase debe estar cada elemento, utilizando como criterio de asignación la proximidad (o similitud) de cada elemento a las distintas clases o grupos existentes. Puede utilizarse, por ejemplo, para identificar el origen geográfico de una muestra de miel desconocida, cuando disponemos de varias muestras de cada zona, utilizando los contenidos o concentraciones de distintos tipos de azúcares que han sido determinados analíticamente en todas las muestras.

La construcción de las funciones discriminantes, combinaciones lineales de las variables originales, se realiza buscando aquellas que hacen máxima la separación entre grupos, lo que equivale a minimizar la variabilidad dentro de las clases. Las funciones obtenidas permiten discriminar de la mejor forma posible entre los elementos pertenecientes a las distintas clases o grupos.

La asignación de los elementos a clases se realiza mediante el criterio de Bayes: cada elemento se asigna a la clase para la que es mayor la probabilidad de pertenencia condicionada por los valores que toman las funciones discriminantes.

XIV.3.1. HIPÓTESIS

Se requiere, para que pueda ser utilizado correctamente, la hipótesis de homogeneidad de varianzas, es decir que las matrices de covarianzas de cada clase sean iguales, o que la dispersión sea similar en cada clase para todas las variables. Se emplea la prueba M de Box para contrastar esta hipótesis de homogeneidad.

También es necesario que las funciones discriminantes tengan distribución Normal, aunque en muchas ocasiones el método funciona razonablemente ante un cierto grado de incumplimiento de estas condiciones.

Para comprobar si el Análisis Discriminante es adecuado podemos utilizar la matriz de clasificación, que muestra el porcentaje de casos asignados correctamente en aquellos elementos de la muestra cuya clase es conocida. Es necesario que el porcentaje de acierto sea alto no solamente en el conjunto de los elementos, sino también en cada uno de los grupos.

XIV.3.2. MÉTODO PASO A PASO

Se puede utilizar para el análisis discriminante el método paso a paso, con el cual se seleccionan sucesivamente de forma automática las variables con mayor capacidad de discriminación de acuerdo con un determinado criterio estadístico (por ejemplo el mínimo valor del coeficiente lambda de Wilks), las cuales se incorporan al conjunto de variables que intervienen en el análisis. En algunos pasos se puede descartar una variable que se introdujo con anterioridad, pero que al considerar otras variables nuevas deja de aportar al conjunto una capacidad de discriminación significativa. En general es preferible sin embargo seleccionar las variables que deben ser utilizadas no de forma automática, sino con criterios técnicos o científicos -no estadísticos- empleando aquellas que

tienen a priori una relación directa (por ejemplo de tipo causa-efecto) con la clasificación.

XIV.3.3. INFLUENCIA DE LAS DISTINTAS VARIABLES INDEPENDIENTES

Generalmente se utiliza la solución obtenida para clasificar casos cuyo grupo es desconocido a priori, es decir con fines de identificación o predicción. También se utiliza para reclasificar los elementos o reasignar los elementos a las distintas clases cuando esa clasificación no es fiable o es solamente orientativa, y para estudiar el efecto de las diferentes variables explicativas sobre la clasificación. Con este último objetivo se utilizan los coeficientes estándar de las primeras funciones discriminantes: cuanto mayor es el coeficiente, mayor es el efecto de esa variable.

También se pueden emplear los llamados "coeficientes de estructura", o coeficientes de correlación entre las funciones discriminantes y las variables independientes (las cuales han sido utilizadas para su construcción). Generalmente los coeficientes de estructura mayores en valor absoluto se corresponden con los mayores coeficientes estándar, aunque en ocasiones el coeficiente de estructura alto corresponde a un coeficiente estándar bajo, lo que permite detectar variables redundantes, o al revés, un coeficiente de estructura bajo con un coeficiente estándar alto, lo que identifica una variable confusora, que interactúa con las restantes distorsionando los resultados. Generalmente ambos tipos de variables deben ser excluidos del análisis.

Se suelen utilizar ANOVAS univariados, para contrastar la hipótesis de que las medias de cada variable independiente son iguales para todos los grupos. Permiten medir también la importancia o capacidad de discriminación de las distintas variables. Si las medias son iguales la variable tendrá en general escasa o nula capacidad de discriminación.

XIV.3.4. VALIDACIÓN

Cuando el número de variables es grande en relación con el tamaño de la muestra, puede aparecer un problema de sobreajuste. Las funciones discriminantes permiten clasificar con un elevado grado de acierto los elementos de la muestra, pero si aplicamos esas mismas funciones a la clasificación de elementos nuevos, no utilizados para la construcción de las funciones discriminantes, el grado de acierto disminuye notablemente. En esos supuestos se están utilizando las peculiaridades de la muestra para asignar correctamente un número alto de casos, pero las funciones no son válidas –no son representativas– para la población de la que procede la muestra. Es necesario realizar algún tipo de validación para asegurar que el procedimiento será válido para la población y no solamente para la muestra.

Una opción es reservar una parte de la muestra (por ejemplo 10 elementos), que no se utilizan para construir las funciones discriminantes, pero son asignados con ellas a las distintas clases. Si la proporción de elementos correctamente clasificados con esta muestra de validación es también elevado, el análisis discriminante será adecuado.

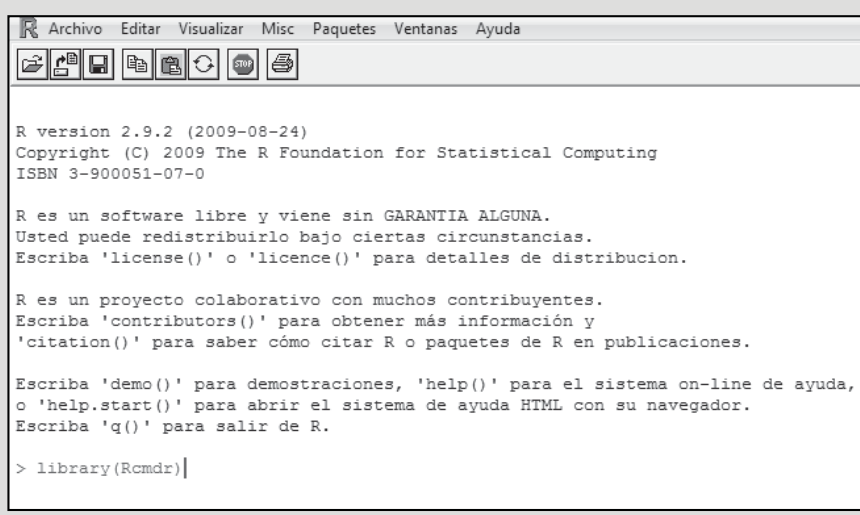
Otra posibilidad es la llamada clasificación cruzada dejando uno fuera. Se realiza un análisis discriminante excluyendo el primer elemento de la muestra, y a continuación este elemento es asignado mediante las funciones discriminantes obtenidas. Posteriormente hacemos lo mismo con cada uno de los elementos de la muestra. De este modo asignamos todos los elementos de la muestra sin que el elemento correspondiente haya sido utilizado en los cálculos de las funciones discriminantes y, por lo tanto, hemos utilizado todos los datos como muestra de validación. Si el porcentaje de acierto con esta validación cruzada es alto y similar, o solo ligeramente inferior, al porcentaje de acierto del análisis discriminante con toda la muestra, la validación es positiva, y el análisis discriminante adecuado para esos datos.

CUADRO XIV.3. Análisis discriminante

Aplicación con R.

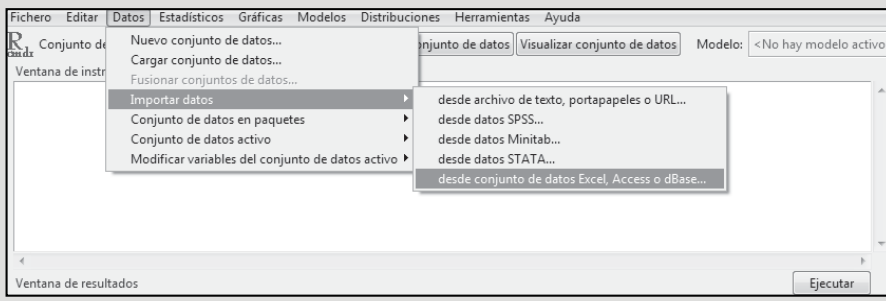
EJEMPLO. En el archivo **Cuadro XIV.3.xls** hay datos obtenidos en un citómetro de flujo del tamaño, rugosidad y emisión de radiación a diferentes longitudes de onda, de células que pertenecen a diferentes especies del fitoplancton: *Rhodomonas baltica*, *Alexandrium tamarense*, y dos cepas de *Alexandrium minutum*. El objetivo es determinar si con las características de las células obtenidas en el citómetro de flujo es posible discriminar entre las diferentes especies, y las cepas dentro de la misma especie.

Paso 1. Después de haber invocado al programa R y seleccionado el directorio, como se explica en el Apéndice II, cargamos el paquete "Rcmdr" con la instrucción library(Rcmdr). Si el paquete no está instalado, la instalación se realiza como se explica en el Apéndice II.

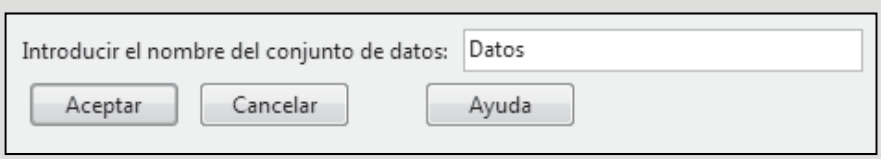

Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> library(Rcmdr)

CUADRO XIV.3. *(Continuación)*[R]

Paso 2. Se selecciona «Datos» ⇨ «Importar datos» ⇨ «desde conjunto de datos Excel, Access o dBase».


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda
Nuevo conjunto de datos...
Cargar conjunto de datos...
Fusionar conjuntos de datos...
Importar datos
Conjunto de datos en paquetes
Conjunto de datos activo
Modificar variables del conjunto de datos activo
desde archivo de texto, portapapeles o URL...
desde datos SPSS...
desde datos Minitab...
desde datos STATA...
desde conjunto de datos Excel, Access o dBase...
Visualizar conjunto de datos
Modelo:
<No hay modelo activo>
Ventana de resultados
Ejecutar

Paso 3. Aparece un pequeño cuadro de diálogo para introducir el nombre del conjunto de datos, en el que dejamos el nombre por defecto «Datos» y pulsamos «Aceptar». Posteriormente aparecen los archivos del directorio de trabajo, en el que elegimos **Cuadro XIV.3.xls** y pulsamos «Abrir».

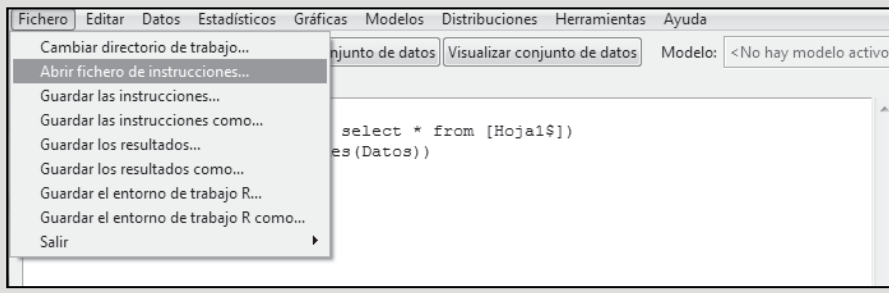


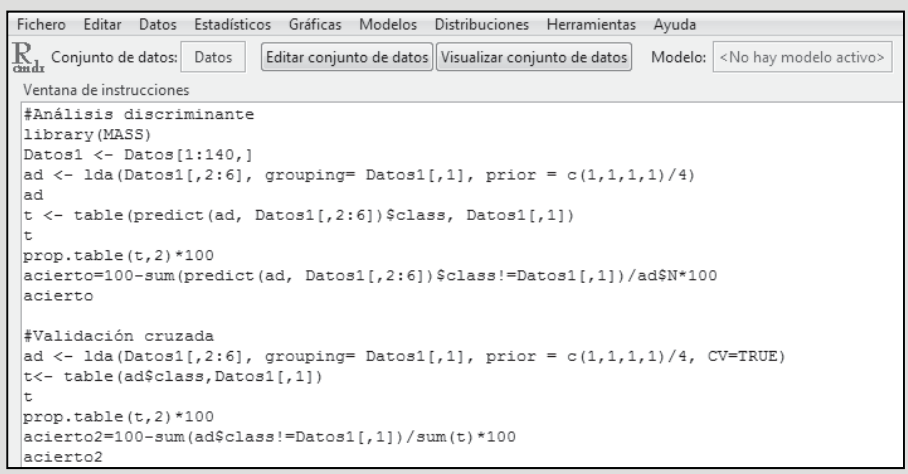
Paso 4. En R Commander pulsamos en «Visualizar conjunto de datos». Es importante inspeccionar los datos y comprobar que no existan líneas, lo cual a veces ocurre al final del archivo, que tengan celdas con el código <NA>. Si existen estas líneas, hay que eliminarlas.

	Especie	Rugosidad	Tamaño	F.515.550	F.560.600	F.650.700	Grupo
1	A. minutum-1	999	921	263	263	208	<NA>
2	A. minutum-1	860	889	267	279	238	<NA>
3	A. minutum-1	996	845	288	279	179	<NA>
4	A. minutum-1	948	821	290	309	474	<NA>
5	A. minutum-1	977	898	293	346	454	<NA>
6	A. minutum-1	829	767	303	299	181	<NA>
7	A. minutum-1	962	770	315	347	330	<NA>
8	A. minutum-1	982	685	336	339	206	<NA>
9	A. minutum-1	1001	751	341	344	229	<NA>
10	A. minutum-1	469	858	343	355	674	<NA>
11	A. minutum-2	844	854	127	14	65	<NA>
12	A. minutum-2	907	588	130	15	37	<NA>
13	A. minutum-2	803	492	162	15	27	<NA>
14	A. minutum-2	882	683	184	15	59	<NA>
15	A. minutum-2	844	698	173	16	25	<NA>
16	A. minutum-2	867	594	127	17	9	<NA>
17	A. minutum-2	600	125	176	17	13	<NA>

CUADRO XIV.3. *(Continuación)*^{R}

Paso 5. A continuación se carga el archivo de instrucciones **Cuadro XIV.3.R** seleccionando «Fichero» ⇨ «Abrir fichero de instrucciones…» y no se guarda la salida.





La primera función carga la librería «MASS»

library(MASS)

Asignamos una variable nueva, «Datos1» al conjunto de los 140 primeros casos, aquellos cuya clase es conocida. Los casos cuya especie es desconocida se ponen al final de la matriz, después del caso 140.

Datos1 <- Datos[1:140,]

A continuación aplicamos la función «lda», que realiza el análisis discriminante lineal. Se indica que las variables independientes están en las columnas 2 a 6 de nuestros datos «Datos1[,2:6]», y la clase o grupo en la columna 1 «Datos1[,1]». Los valores 2 y 6 y la posición de la columna que informa de la especie a la que pertenece cada caso se debe cambiar en todos los sitios que aparezca en el archivo de instrucciones, para una aplicación con otros datos.

CUADRO XIV.3. *(Continuación)*[R]

La columna 7, que tiene la variable llamada «Grupo», es simplemente para probar la capacidad predictiva del análisis, ya que indica a que especie pertenecen los casos que no se han empleado en el análisis y que están al final de la matriz de datos.

El tercer argumento «prior» indica que las probabilidades a priori son iguales (1/4) para todas las clases. Si tuviésemos dos clases deberíamos escribir «prior = c(1,1,)/2». Como hay 4 especies, entonces ponemos «prior = c(1,1,1,1)/4». Si no ponemos nada (si suprimimos este tercer argumento, incluida la coma) entenderá que deseamos que las probabilidades a priori se determinen por el número de elementos que tiene cada clase en la muestra. Generalmente es preferible la opción de probabilidades iguales dado que el número de elementos de cada clase en la muestra no suele ser representativo de la población real. Las probabilidades a priori se utilizan mediante la regla de Bayes para asignar los elementos a clases.

```
ad <- lda(Datos1[,2:6], grouping= Datos1[,1], prior = c(1,1,1,1)/4)
```

La siguiente función es para escribir los resultados

```
ad
```

El mejor método para comprobar la capacidad predictiva del análisis es la validación cruzada, o clasificación "dejando uno fuera".

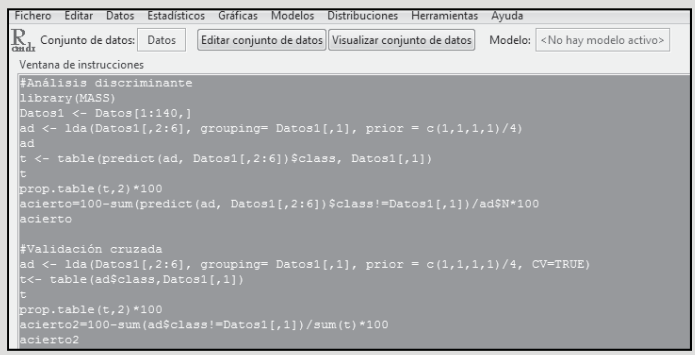
```
ad <- lda(Datos1[,2:6], grouping= Datos1[,1], prior = c(1,1,1,1)/4, CV=TRUE)
t<- table(ad$class,Datos1[,1])
t
prop.table(t,2)*100
acierto2=100-sum(ad$class!=Datos1[,1])/sum(t)*100
acierto2
```

Por último, también es posible validar la capacidad predictiva del modelo utilizando una muestra separada, que no haya sido empleada en el cálculo de la función discriminante. Recordemos que en nuestra muestra había 8 casos –situados en último lugar– para los cuales no se indica a que especie pertenecen en la columna 1 y, por ello, no se usan en el análisis. En general utilizaremos el análisis discriminante para identificar las muestras desconocidas, pero como en nuestro ejemplo si conocemos a que especies pertenecen estos 8 elementos, podemos utilizarlos como muestra de validación.

```
Datos2 <- Datos[141:148,]
z <- lda(Datos1[,2:6], grouping= Datos1[,1], prior = c(1,1,1,1)/4)
predict(z, Datos2[,2:6])$class
```

CUADRO XIV.3. *(Continuación)*[R]

Paso 6. Por último, se seleccionan todas las instrucciones del paso 5 y se pulsa «Ejecutar».



Paso 7. Los resultados obtenidos son:

a) Medias de grupo. Muestran el valor medio de cada variable para cada una de las especies.

```
Group means:
             Rugosidad   Tamaño F.515.550 F.560.600 F.650.700
A. minutum-1  932.7647 826.9412 257.05882  275.1765  301.2353
A. minutum-2  850.6400 654.7200 218.52000   74.4400   49.5600
A. tamarense  822.4286 793.7143 149.52381  179.3571  842.3333
R. baltica    859.7679 593.8393  36.92857  166.1786  290.3929
```

b) Coeficientes de la función discriminante (siempre hay una función menos que el número de clases, salvo que el número de variables sea más pequeño). Estos coeficientes muestran la influencia de cada variable en la función, aunque están condicionados por los valores medios que toma cada variable y no deben ser interpretados directamente con este fin, ya que no son coeficientes estandarizados.

```
Coefficients of linear discriminants:
                     LD1           LD2           LD3
Rugosidad  0.0022006153 -0.0007762740 -0.0006119396
Tamaño    -0.0009839004 -0.0002579719 -0.0056520454
F.515.550 -0.0022125341  0.0288049125  0.0049028945
F.560.600 -0.0066771910 -0.0096999000 -0.0157301799
F.650.700  0.0174729045  0.0029254158  0.0021428518
```

c) A continuación se observa el porcentaje de varianza explicado por cada función discriminante. El primer eje absorbe un 78,58% y el segundo eje el 16,27%. Como los dos primeros ejes recogen el 94,85% de la varianza, el gráfico bipolar con los dos primeros ejes discriminantes sería suficiente para poder tener una representación de los datos.

```
Proportion of trace:
   LD1    LD2    LD3
0.7858 0.1627 0.0515
```

CUADRO XIV.3. *(Continuación)*[R]

d) La matriz de clasificación muestra que el 100% de los casos son asignados correctamente.

```
              A. minutum-1 A. minutum-2 A. tamarense R. baltica
  A. minutum-1           17            0            0          0
  A. minutum-2            0           25            0          0
  A. tamarense            0            0           42          0
  R. baltica              0            0            0         56

> prop.table(t,2)*100

              A. minutum-1 A. minutum-2 A. tamarense R. baltica
  A. minutum-1          100            0            0          0
  A. minutum-2            0          100            0          0
  A. tamarense            0            0          100          0
  R. baltica              0            0            0        100

> acierto=100-sum(predict(ad, Datos1[,2:6])$class!=Datos1[,1])/ad$N*100

> acierto
[1] 100
```

e) La validación cruzada muestra también un alto grado de casos asignados correctamente (99,28%), de tal forma que solo un caso de *A. minutum*-1 se identifica como *A. tamarense*.

```
              A. minutum-1 A. minutum-2 A. tamarense R. baltica
  A. minutum-1           16            0            0          0
  A. minutum-2            0           25            0          0
  A. tamarense            1            0           42          0
  R. baltica              0            0            0         56

> prop.table(t,2)*100

              A. minutum-1 A. minutum-2 A. tamarense R. baltica
  A. minutum-1    94.117647     0.000000     0.000000   0.000000
  A. minutum-2     0.000000   100.000000     0.000000   0.000000
  A. tamarense     5.882353     0.000000   100.000000   0.000000
  R. baltica       0.000000     0.000000     0.000000 100.000000

> acierto2=100-sum(ad$class!=Datos1[,1])/sum(t)*100

> acierto2
[1] 99.28571
```

f) Los últimos 6 casos, aunque se conocía a que especie pertenecían, se usaron para comprobar la capacidad de predicción del análisis. Se observa que identifica correctamente el 100% de los casos.

```
> predict(z, Datos2[,2:6])$class
[1] A. minutum-1 A. minutum-1 A. minutum-2 A. minutum-2 A. tamarense A. tamarense R. baltica   R. baltica
Levels: A. minutum-1 A. minutum-2 A. tamarense R. baltica
```

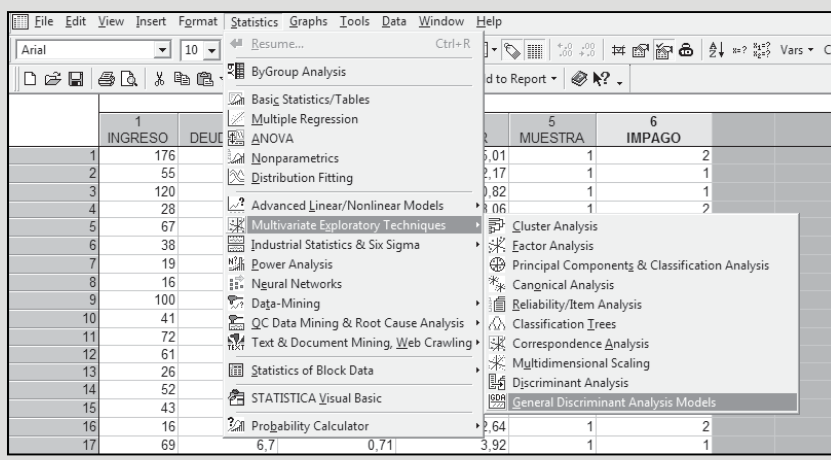
Especie	Rugosidad	Tamaño	F.515.550	F.560.600	F.650.700	Grupo
	914	790	262	277	244	*A. minutum*-1
	878	863	263	276	233	*A. minutum*-1
	882	534	166	19	228	*A. minutum*-2
	891	599	172	19	55	*A. minutum*-2
	842	784	160	209	864	*A. tamarense*
	873	783	186	209	865	*A. tamarense*
	877	635	23	298	434	*R. baltica*
	886	642	28	305	476	*R. baltica*

CUADRO XIV.3. *(Continuación)*[STATISTICA]

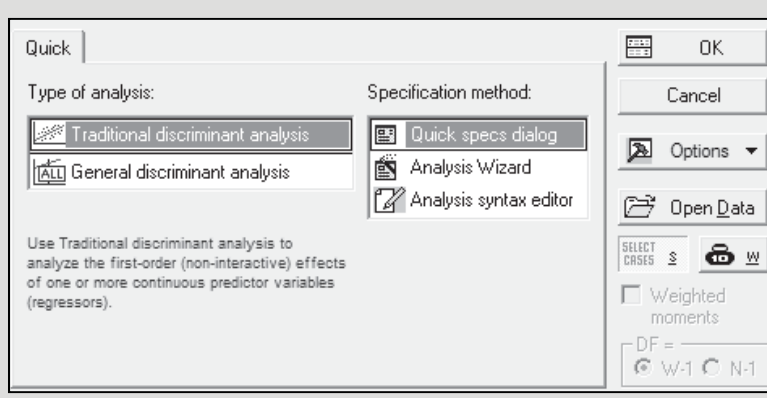
Aplicación con STATISTICA.

EJEMPLO. Utilizaremos los datos del archivo **Cuadro XIV.3.sta**, con 100 clientes de un banco de los cuales se tiene información acerca de 4 variables: ingresos, relación entre deudas totales e ingresos (en porcentaje), deudas mediante tarjeta de crédito, y otras deudas. La última variable «Impago» nos indica si el cliente ha tenido impagos (2) o no (1). La variable «Muestra» toma el valor 1 para los 70 primeros casos, y 0 para los 30 últimos, que utilizaremos como muestra de validación. Con el análisis discriminante se pretende averiguar si existe relación entre los impagos y el resto de las variables, y por lo tanto si el banco será capaz de predecir el impago de un cliente conociendo los valores de las otras variables.

Paso 1. Se selecciona «Statistics» ⇨ «Multivariate Exploratory Techniques» ⇨ «General Discriminant Analysis Models».

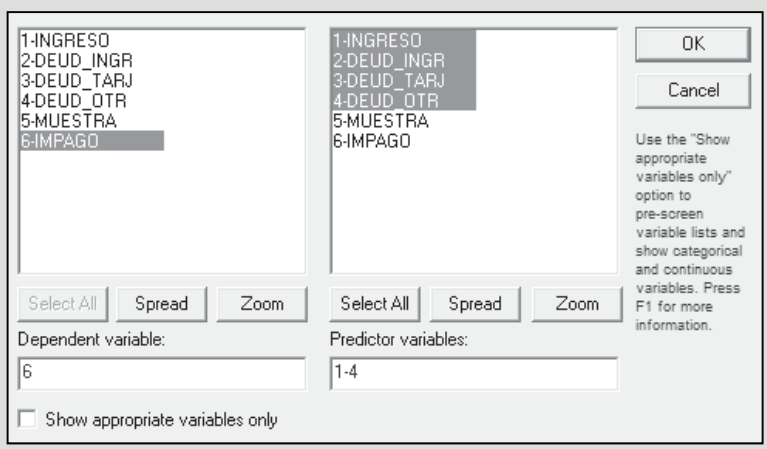


Paso 2. En la ventana que aparece seleccionamos «Traditional discriminant analysis» y «Quick specs dialog».

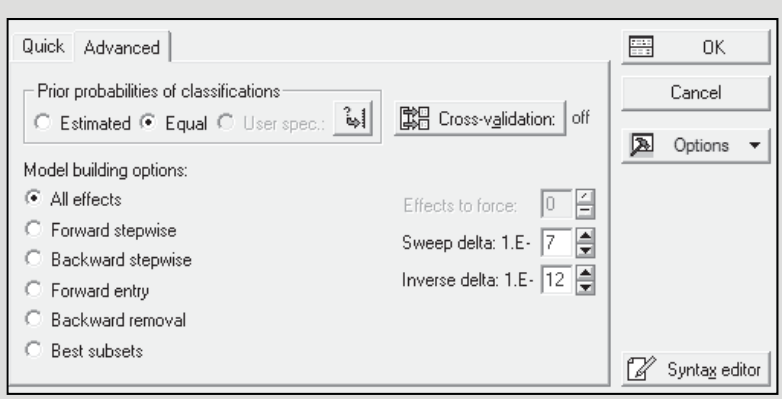


CUADRO XIV.3. *(Continuación)*[STATISTICA]

Paso 3. En la pestaña «Quick» pulsamos en «Variables» y elegimos las cuatro primeras en la lista de variables independientes o «Predictor variables: », y la sexta en «Dependent variable:» o variable de clasificación.

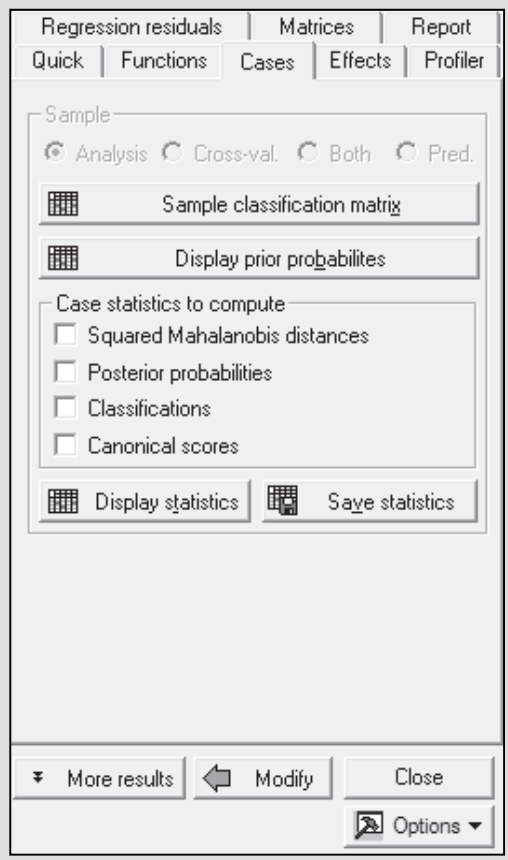


Paso 4. En la pestaña «Advanced» marcamos la opción de probabilidades a priori iguales «Equal».



CUADRO XIV.3. *(Continuación)*[STATISTICA]

Paso 5. En la ventana de resultados elegimos en primer lugar la pestaña «Cases», y dentro de ella la opción «Sample clasification matrix».

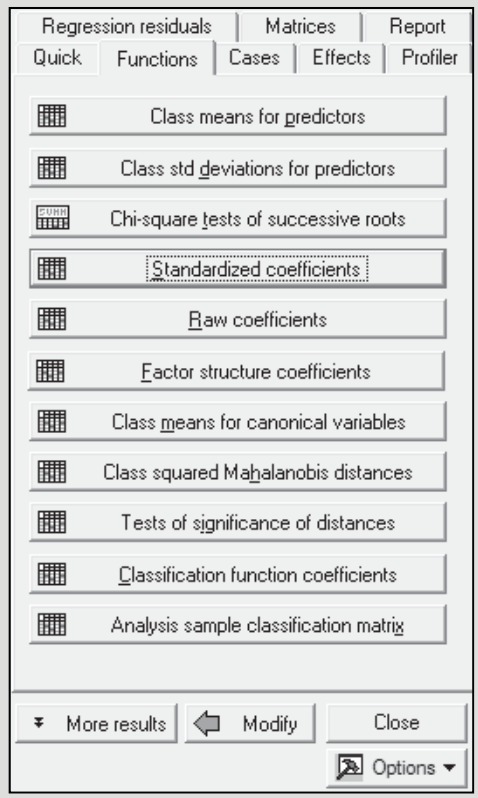


Paso 6. El resultado nos muestra la tabla de clasificación. En filas la clase original, y en columnas la clase de predicción utilizando la función discriminante. Podemos observar como todos los casos han sido identificados correctamente, 100% de acierto y, por lo tanto, el análisis discriminante permite pronosticar, en función de los valores de las variables explicativas, si habrá o no impagos.

	Classification Matrix (Cuadro XIV.3.sta) Rows: Observed classifications Columns: Predicted classifications		
Class	Percent Correct	1 p=,5000	2 p=,5000
1	100,0000	68,00000	0,00000
2	100,0000	0,00000	32,00000
Total	100,0000	68,00000	32,00000

CUADRO XIV.3. *(Continuación)* *STATISTICA*

Paso 7. Volvemos a la ventana de resultados y pulsamos en la pestaña «Functions». Dentro de ella elegimos «Standardized coefficients» (coeficientes estándar de la función discriminante).



Paso 8. Estos valores muestran la distinta influencia que tiene cada variable en el proceso de clasificación. La relación entre deudas e ingresos y otras deudas son las dos variables que tienen mayor efecto, y el nivel de ingresos es la menos importante. La tabla también incluye los «Eigenvalue» (autovalores), que indican la variabilidad recogida por cada una de las funciones discriminantes, y el porcentaje correspondiente a cada una «Cum. sProp.»; como hay una única función, ésta absorbe el 100% de la varianza.

	Standardized Canonical Discriminant Function Coefficients (Cuadro XIV.3.sta) Sigma-restricted parameterization
Effect	Function 1
Intercept	0,00000
INGRESO	0,27941
DEUD_INGR	-1,57249
DEUD_TARJ	-0,88238
DEUD_OTR	1,51821
Eigenvalue	3,62664
Cum. Prop.	1,00000

CUADRO XIV.3. *(Continuación)*[STATISTICA]

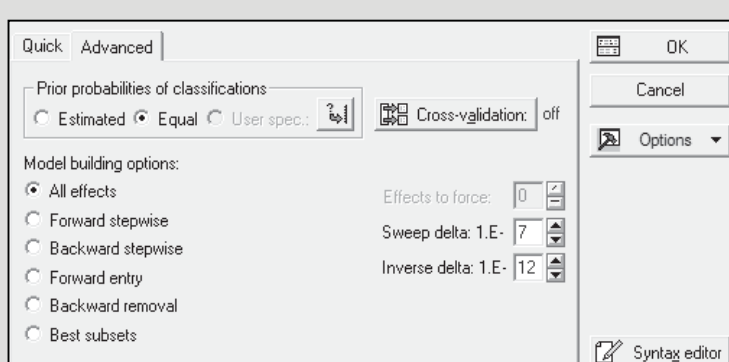
Paso 9. Con el mismo fin inspeccionamos los coeficientes estructura, «Factor structure coefficients» (o coeficientes de correlación entre las variables originales y la función discriminante). Estos coeficientes indican con qué función está más correlacionada cada variable y, por tanto, indican cuales son las variables más importantes dentro de cada función. El valor mayor corresponde también a la relación entre deudas e ingresos «DEUD_INGR», y el nivel de ingresos tiene un coeficiente de estructura bajo. También la variable otras deudas «DEUDA_OTR» tiene un coeficiente de estructura bajo, lo que contrasta con su elevado coeficiente estándar, que vimos en la tabla anterior. Ello indica un cierto carácter de variable confusora en otras deudas. El elevado porcentaje de acierto desaconseja sin embargo repetir el análisis prescindiendo de esta variable.

	Factor Structure Coefficients (Cuadro XIV.3.sta) Sigma-restricted parameterization
	Function
Effect	1
Intercept	
INGRESO	0,171651
DEUD_INGR	-0,560912
DEUD_TARJ	-0,228595
DEUD_OTR	-0,086744

Paso 10. En la ventana del paso 7 pulsamos en «Chi-square of successive roots». El estadístico Lambda se utiliza para seleccionar las funciones discriminantes significativas. En este ejemplo se observa que la única función que tenemos es significativa, ya que $p < 0,05$.

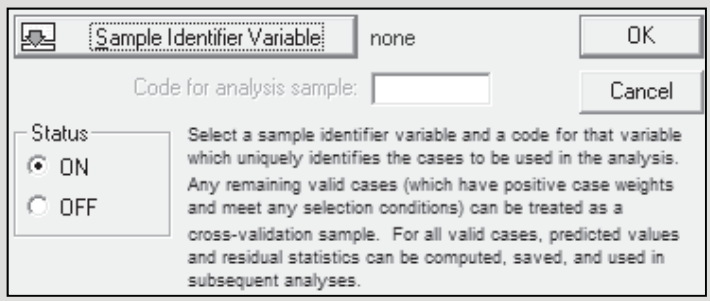
	Chi-Square Tests with Successive Roots Removed (Cuadro XIV.3.sta) Sigma-restricted parameterization					
Removed	Eigen-value	Canonicl R	Wilk's Lambda	Chi-Sqr.	df	p-level
0	3,626644	0,885359	0,216139	147,0559	4,000000	0,00

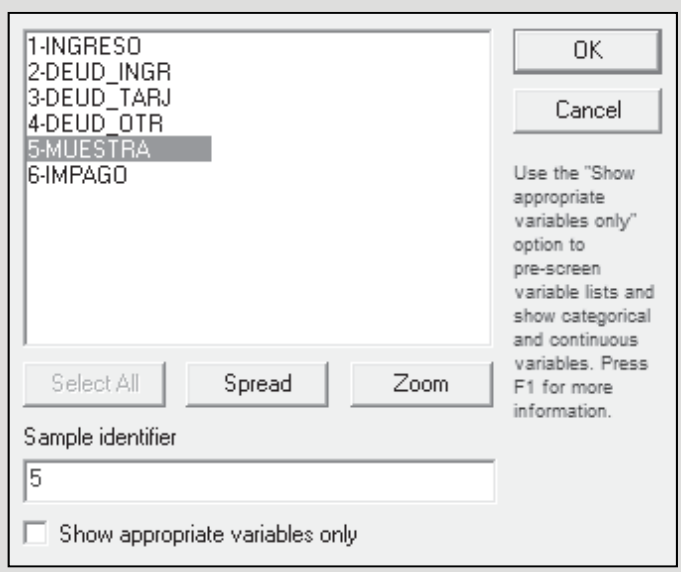
Paso 11. Es necesario validar los resultados mediante una muestra de contraste. Para ello, repetimos los pasos 1 a 4 y en la pestaña «Advanced» marcamos la opción de probabilidades a priori iguales «Equal» y pulsamos «Cross-validation».



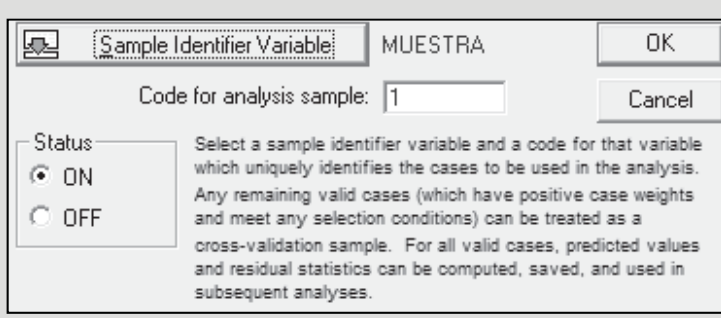
CUADRO XIV.3. *(Continuación)*[STATISTICA]

Paso 12. Marcamos el status «ON» y pulsamos en «Sample identifier Variable», para indicar cuál es la muestra de análisis y cual la de validación. Utilizaremos los primeros 70 casos (identificados con la variable «Muestra» = 1) para realizar el análisis, y los 30 siguientes para el contraste.


Sample Identifier Variable
none
OK
Code for analysis sample:
Cancel
Status
ON
OFF
Select a sample identifier variable and a code for that variable which uniquely identifies the cases to be used in the analysis. Any remaining valid cases (which have positive case weights and meet any selection conditions) can be treated as a cross-validation sample. For all valid cases, predicted values and residual statistics can be computed, saved, and used in subsequent analyses.

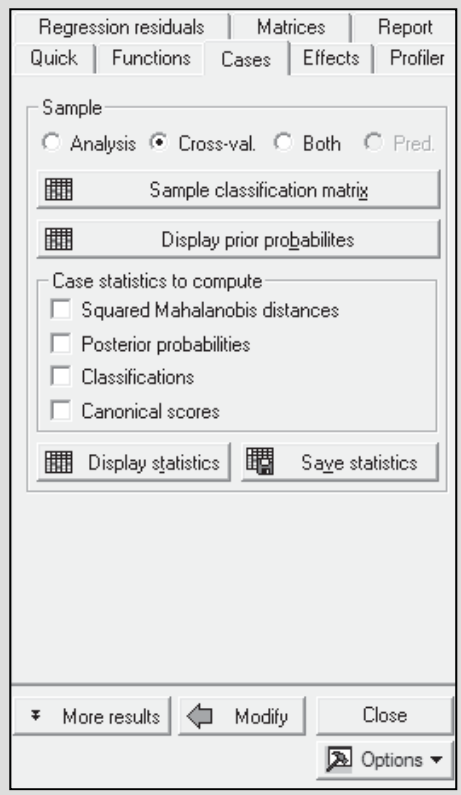


Debe aparecer el valor del código seleccionado «1».


Sample Identifier Variable
MUESTRA
OK
Code for analysis sample:
1
Cancel
Status
ON
OFF
Select a sample identifier variable and a code for that variable which uniquely identifies the cases to be used in the analysis. Any remaining valid cases (which have positive case weights and meet any selection conditions) can be treated as a cross-validation sample. For all valid cases, predicted values and residual statistics can be computed, saved, and used in subsequent analyses.

CUADRO XIV.3. *(Continuación)*[STATISTICA]

Paso 13. En la ventana de resultados elegimos la pestaña «Cases», marcamos la opción «Cross-val» y pulsamos «Sample clasification matrix».



Paso 14. La matriz de clasificación que aparece contiene únicamente los 30 elementos de la muestra de validación, que han sido asignados mediante la función discriminante calculada con los 70 elementos de la muestra de análisis. Todos los elementos son asignados correctamente, de nuevo 100% de acierto y, por lo tanto, la validación es positiva: el procedimiento puede ser utilizado para pronosticar el impago con clientes nuevos.

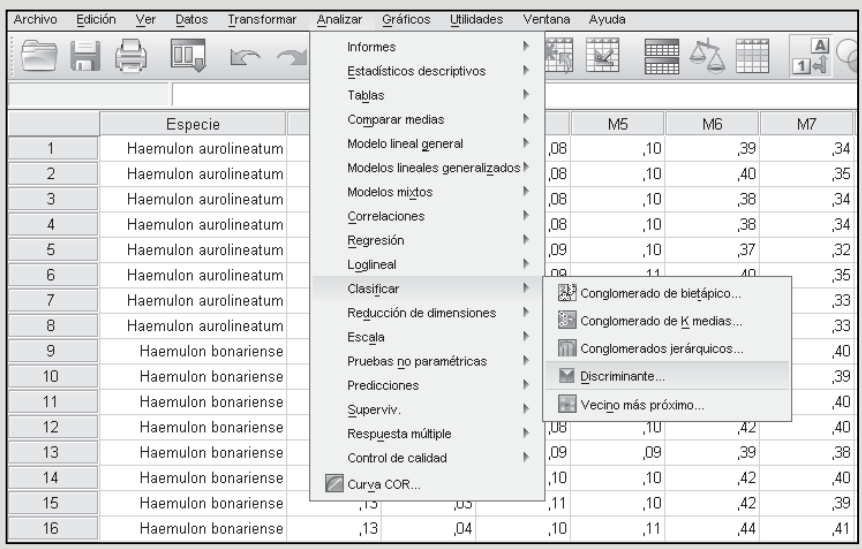
	Classification Matrix (Cuadro XIV.3.sta) Rows: Observed classifications Columns: Predicted classifications		
Class	Percent Correct	1 p=,5000	2 p=,5000
1	100,0000	20,00000	0,00000
2	100,0000	0,00000	10,00000
Total	100,0000	20,00000	10,00000

CUADRO XIV.3. *(Continuación)*[SPSS]

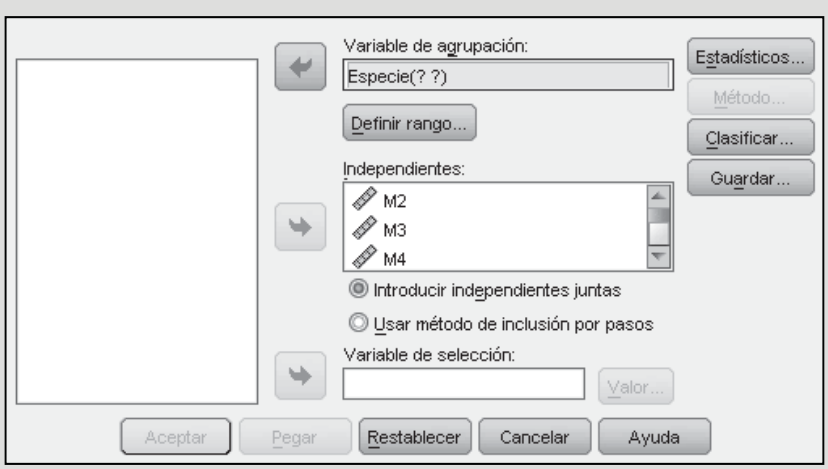
Aplicación con SPSS.

EJEMPLO. En el archivo **Cuadro XIV.3.sav** hay datos de medidas del cuerpo, estandarizadas en función de la longitud estándar, de 7 especies de peces del género Haemulon. El objetivo es determinar si en función de las medidas biométricas es posible discriminar entre las especies.

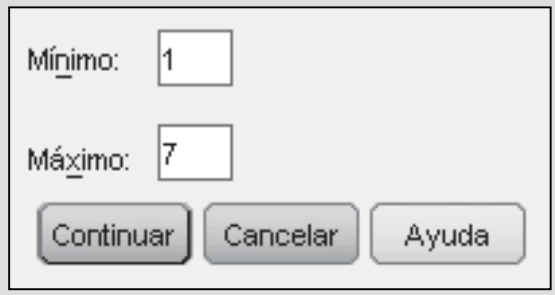
Paso 1. Se selecciona «Analizar» ⇨ «Clasificar» ⇨ «Discriminante…».



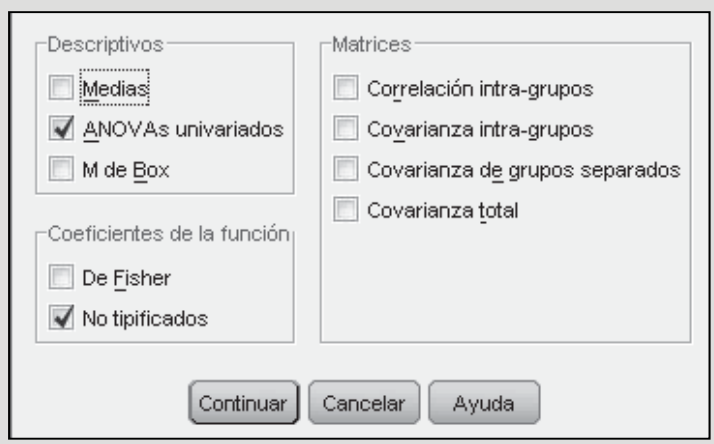
Paso 2. En la ventana del análisis discriminante introducimos las variables independientes, que son todas las medidas, y la variable de agrupación «Especie», la cual tiene que ser numérica. A continuación, mediante «Definir rango» establecemos los valores mínimo y máximo de los códigos de las clases que queremos utilizar en el análisis, en este ejemplo desde 1 hasta 7 especies.



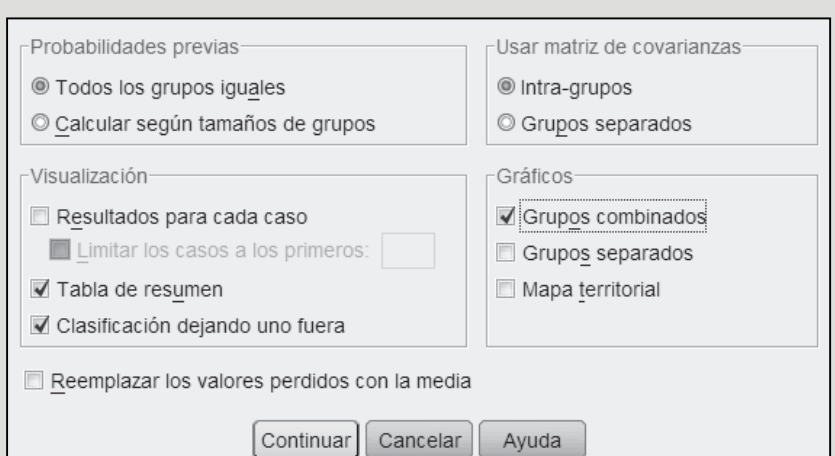
CUADRO XIV.3. *(Continuación)*[SPSS]



Paso 3. En la ventana del paso 2 pulsamos «Estadísticos» y marcamos las opciones «ANOVAs univariados» y coeficientes «No tipificados».



Paso 4. En la ventana del paso 2 pulsamos «Clasificar» y marcamos las opciones «Todos los grupos iguales» en «Probabilidades previas». En «Visualización» se marca «Tabla de resumen» y «Clasificación dejando uno fuera». Por último, marcamos «Grupos combinados» en «Gráficos».



CUADRO XIV.3. *(Continuación)*[SPSS]

Paso 5. En primer lugar observamos una prueba de igualdad de medias, o ANOVA, para cada variable, que trata de establecer si la media es distinta en las diferentes clases (al menos en dos de ellas). Los valores *F* más grandes indican las variables que podrían discriminar mejor entre las clases, que son en este caso M2, M11, M3, etc. En todo caso esta prueba no tiene en cuenta la interacción entre variables, de gran importancia en el análisis discriminante, y no se comprueban las hipótesis de este test, por tanto, la información es sólo orientativa.

Pruebas de igualdad de las medias de los grupos

	Lambda de Wilks	F	gl1	gl2	Sig.
M2	,157	62,854	6	70	,000
M3	,238	37,393	6	70	,000
M4	,269	31,675	6	70	,000
M5	,327	23,988	6	70	,000
M6	,508	11,302	6	70	,000
M7	,411	16,709	6	70	,000
M8	,565	8,977	6	70	,000
M9	,646	6,386	6	70	,000
M10	,694	5,146	6	70	,000
M11	,222	40,950	6	70	,000
M12	,761	3,655	6	70	,003
M13	,256	33,846	6	70	,000
M14	,654	6,173	6	70	,000
M15	,658	6,076	6	70	,000
M16	,619	7,182	6	70	,000
M19	,767	3,539	6	70	,004
M20	,637	6,642	6	70	,000
M21	,629	6,895	6	70	,000
M22	,359	20,799	6	70	,000
M23	,352	21,458	6	70	,000
M24	,745	3,983	6	70	,002
M25	,266	32,222	6	70	,000
M26	,414	16,481	6	70	,000
M27	,596	7,894	6	70	,000
M30	,314	25,542	6	70	,000
M31	,722	4,482	6	70	,001
M32	,637	6,637	6	70	,000

Paso 6. Las funciones discriminantes son unas expresiones matemáticas que nos permiten ver la separación de los grupos. Lo que nos indica la siguiente tabla es el porcentaje de la varianza de nuestros datos explicado por estas funciones discriminantes. En el ejemplo tenemos 6 funciones (7 grupos menos 1), de las cuales la primera explica el 48,5% de la varianza, la segunda el 23%, etc. En este caso entre las dos primeras funciones discriminantes se explica sólo el 71,4 % de la varianza, por lo que para la representación gráfica y la interpretación de los datos, no sería suficiente con la representación de las dos primeras funciones. Con las coordenadas polares (Capítulo I) se podrían representar varias funciones. La tabla también incluye los «Autovalores», que son las varianzas explicadas por cada una de las funciones.

CUADRO XIV.3. *(Continuación)*[SPSS]

Autovalores

Función		Autovalor	% de varianza	% acumulado	Correlación canónica
0	1	26,526[a]	48,5	48,5	,982
	2	12,570[a]	23,0	71,4	,962
	3	7,209[a]	13,2	84,6	,937
	4	4,438[a]	8,1	92,7	,903
	5	2,240[a]	4,1	96,8	,831
	6	1,757[a]	3,2	100,0	,798

a. Se han empleado las 6 primeras funciones discriminantes canónicas en el análisis.

Paso 7. El estadístico Lambda se utiliza para seleccionar las funciones discriminantes significativas. En este ejemplo se observa que todas las funciones son significativas, ya que $p < 0,05$.

Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1 a la 6	,000	702,752	162	,000
2 a la 6	,000	507,160	130	,000
3 a la 6	,003	353,297	100	,000
4 a la 6	,021	229,087	72	,000
5 a la 6	,112	129,178	46	,000
6	,363	59,828	22	,000

Paso 8. Los coeficientes estandarizados permiten medir la influencia de las distintas variables. Por ejemplo, en la primera función discriminante, las variables que más influyen son M2, M32, M6 y M7.

Coeficientes estandarizados de las funciones discriminantes canónicas

	Función					
	1	2	3	4	5	6
M2	-1,174	-,010	,673	,080	-,270	-,273
M3	-,779	-,276	,182	,062	-,099	-,191
M4	,380	,195	-,058	-,160	,011	,202
M5	,471	,087	,554	,409	,201	-,440
M6	-,955	-,359	-,650	-,191	-,612	-,147
M7	,839	,553	-,015	-,191	,876	,665
M8	,562	-,279	,181	-,470	,702	-,394
M9	-,355	-,230	-,215	,439	,416	,344
M10	,178	,433	,034	,365	-,811	,476
M11	-,138	,804	-,018	-,385	-,426	-,916
M12	-,156	-,623	,123	,578	-,171	,096
M13	-,108	,165	-,244	-,222	,340	,495
M14	,270	,031	,288	-,019	-,040	,411
M15	-,013	-,094	-,305	,319	-,226	,058
M16	-,285	-,106	,383	-,500	-,029	,127
M19	-,332	,093	,008	,389	-,033	-,480
M20	-,156	,065	-,438	-,137	-,404	,058
M21	,107	-,323	,288	-,384	-,485	,570
M22	,363	,084	,258	-,435	-,091	-,061
M23	,117	,166	,348	,131	,218	-,090
M24	,564	-,161	,061	-,225	,630	,100
M25	-,370	,847	,128	,821	-,495	,219
M26	-,432	,249	-,843	,422	,917	-,427
M27	,699	-,221	,076	-,278	-,460	,025
M30	,205	,136	,523	,493	,146	,499
M31	,067	-,231	-,218	,179	,568	,090
M32	1,081	,580	,230	,258	-,176	,132

CUADRO XIV.3. *(Continuación)*[SPSS]

Paso 9. Los coeficientes de estructura son los coeficientes de correlación entre las funciones discriminantes y las variables independientes. Estos coeficientes indican con qué función está más correlacionada cada variable. Por ejemplo, las variables M3, M4 y M5 están más correlacionadas con la función 1 que con el resto de funciones. Las variables M11, M13, M25, M23, M10, M20 y M32 con la función 2, y así sucesivamente. En general, los mayores coeficientes de estructura se corresponden con los mayores coeficientes estándar, y por lo tanto también identifican a las variables más importantes. La discrepancia entre ambos se utiliza para detectar variables redundantes (coeficiente estructura mucho mayor que coeficiente estándar) o confusoras (a la inversa); ambos tipos de variables distorsionan los resultados y en general deben ser suprimidas del análisis.

Matriz de estructura

	Función					
	1	2	3	4	5	6
M3	-,334*	,115	,073	,098	-,035	,033
M4	-,263*	,160	,247	-,029	,192	-,140
M5	,261*	,081	-,032	,170	,044	-,133
M11	-,072	,492*	,003	-,248	,049	-,162
M13	-,113	,425*	-,127	-,104	,139	,219
M25	,121	,381*	-,014	,349	-,060	,054
M23	-,043	,327*	,199	-,121	,197	,051
M10	-,021	,171*	,007	-,062	,136	,033
M20	-,048	,157*	-,134	-,118	-,055	-,048
M32	,090	,136*	,094	,111	-,004	,031
M2	-,380	,066	,380*	-,298	,141	-,119
M30	-,206	-,068	,352*	,027	,152	,176
M26	-,067	,185	-,298*	,099	,286	,006
M15	,006	,072	-,238*	,014	-,026	,156
M22	,091	,288	,014	-,332*	,109	,054
M6	-,111	,084	-,179	-,257*	,033	,122
M7	-,078	,244	-,152	-,254*	,052	,188
M12	-,020	,076	-,121	,157*	,077	,031
M8	-,003	,108	,179	-,125	,327*	-,217
M9	-,004	,107	,154	-,048	,307*	-,083
M16	,062	,160	,041	,021	,266*	,107
M24	,080	,007	,057	-,052	,244*	,031
M31	,060	,124	,062	,046	,160*	,026
M21	,103	,077	,062	-,005	-,114	,321*
M27	,135	-,004	-,030	-,068	-,127	-,269*
M14	-,035	,054	-,179	-,102	-,168	,261*
M19	-,020	,004	,059	,183	-,132	-,213*

Correlaciones intra-grupo combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas

Variables ordenadas por el tamaño de la correlación con la función.

*. Mayor correlación absoluta entre cada variable y cualquier función discriminante.

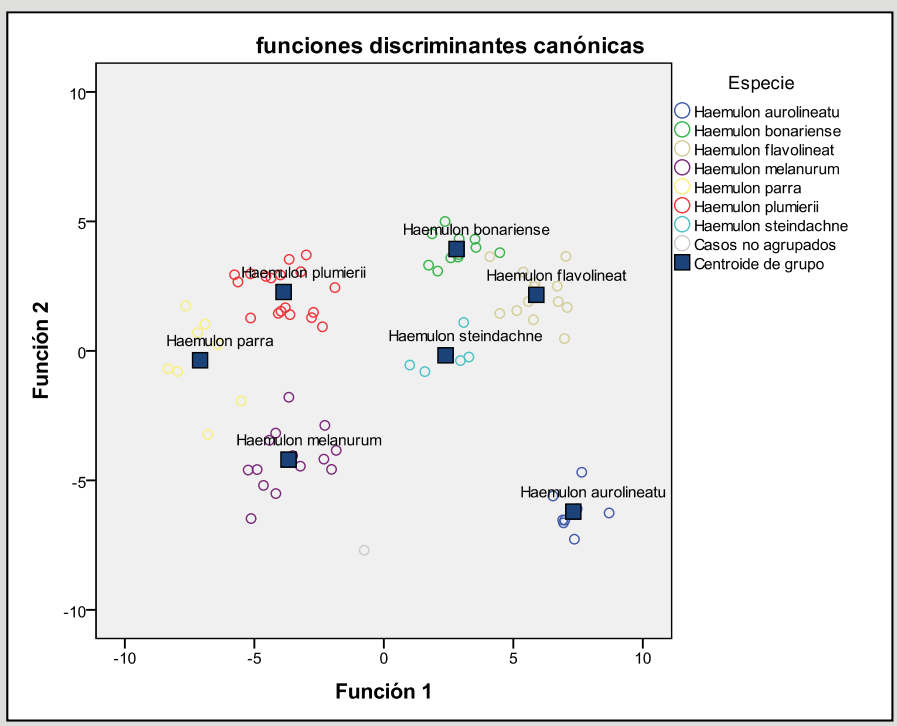
CUADRO XIV.3. *(Continuación)*[SPSS]

Paso 10. Los valores medios «Centroides» de las funciones discriminantes sirven para caracterizar a cada grupo, definiendo su perfil típico. Cuanto más próximos estén a ese perfil los distintos elementos de cada grupo, más clara será su clasificación o identificación.

Funciones en los centroides de los grupos

Especie	Función					
	1	2	3	4	5	6
Haemulon aurolineatum	7,317	-6,207	,704	-2,259	-1,813	-,353
Haemulon bonariense	2,806	3,931	-3,605	-1,840	,019	1,547
Haemulon flavolineatum	5,893	2,159	1,169	3,373	-,103	-,254
Haemulon melanurum	-3,680	-4,199	-,139	1,199	1,318	1,374
Haemulon parra	-7,095	-,364	-3,966	1,374	-2,008	-1,637
Haemulon plumierii	-3,870	2,273	3,348	-1,330	-,318	-,097
Haemulon steindachneri	2,384	-,174	-1,553	-1,875	3,792	-3,059

Paso 11. El gráfico conjunto de los elementos de la muestra en el plano de las dos primeras funciones discriminantes muestra cómo las especies están razonablemente separadas. Este gráfico no es muy orientativo cuando la validación cruzada muestra un porcentaje de predicción bajo.



CUADRO XIV.3. *(Continuación)*[SPSS]

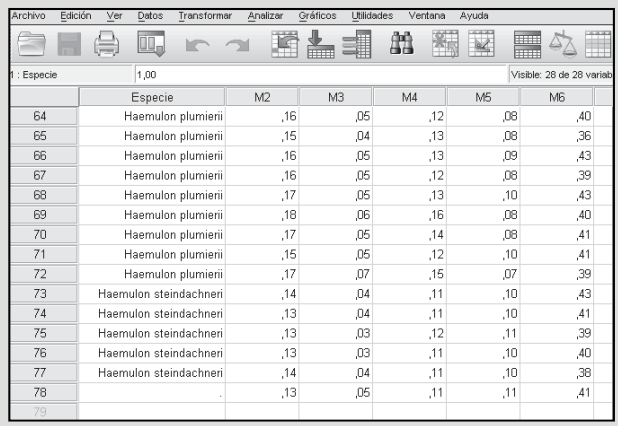
Paso 12. El resultado de mayor interés es la tabla de clasificación, la cual se muestra solo parcialmente con algunas especies.

		Especie	Grupo de pertenencia pronosticado		
			Haemulon aurolineatum	Haemulon bonariense	Haemulon flavolineatum
Original	Recuento	Haemulon aurolineatum	8	0	0
		Haemulon bonariense	0	11	0
		Haemulon flavolineatum	0	0	13
		Haemulon melanurum	0	0	0
		Haemulon parra	0	0	0
		Haemulon plumierii	0	0	0
		Haemulon steindachneri	0	0	0
		Casos desagrupados	0	0	0
	%	Haemulon aurolineatum	100,0	,0	,0
		Haemulon bonariense	,0	100,0	,0
		Haemulon flavolineatum	,0	,0	100,0
		Haemulon melanurum	,0	,0	,0
		Haemulon parra	,0	,0	,0
		Haemulon plumierii	,0	,0	,0
		Haemulon steindachneri	,0	,0	,0
		Casos desagrupados	,0	,0	,0
Validación cruzada[a]	Recuento	Haemulon aurolineatum	8	0	0
		Haemulon bonariense	0	11	0
		Haemulon flavolineatum	0	0	13
		Haemulon melanurum	0	0	0
		Haemulon parra	0	0	0
		Haemulon plumierii	0	1	0
		Haemulon steindachneri	0	1	0
	%	Haemulon aurolineatum	100,0	,0	,0
		Haemulon bonariense	,0	100,0	,0
		Haemulon flavolineatum	,0	,0	100,0
		Haemulon melanurum	,0	,0	,0
		Haemulon parra	,0	,0	,0
		Haemulon plumierii	,0	5,6	,0
		Haemulon steindachneri	,0	20,0	,0

Esta tabla muestra -para cada clase o grupo original- el número de elementos asignados a las diferentes clases. El porcentaje de asignación correcta 100%, es decir, el 100% de los casos se asignan correctamente a su grupo, indica un buen funcionamiento del análisis discriminante. La validación cruzada, en la mitad inferior de la tabla, muestra unos porcentajes de acierto menores que la matriz de clasificación original, con un 97,4% de acierto global que se distribuye de forma variable entre los diferentes grupos. En la especie *Haemulun plumierii* confunde un individuo con *Haemulon bonariense*, y en *Haemulon steindachneri* también confunde un individuo con *Haemulon bonariense*. El resto de especies las identifica correctamente en el 100% de los casos. El porcentaje de la validación cruzada da una mejor idea del grado de discriminación que existe entre las especies, en base a su morfometría, que el porcentaje original de asignación, que en este caso es del 100%.

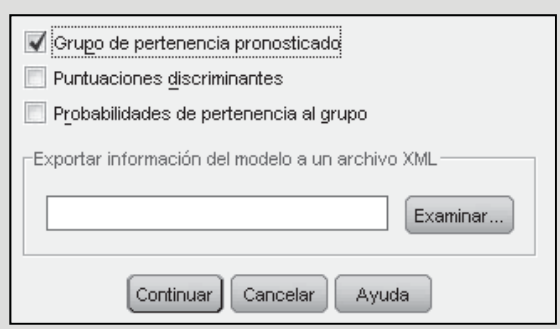
CUADRO XIV.3. *(Continuación)*[SPSS]

Paso 13. Por último, el análisis discriminante permite predecir a qué grupo pertenece un determinado individuo. En el archivo de datos se observa que el caso 78 no tiene asignado grupo. El porcentaje de acierto de la validación cruzada (97,4%), es el grado de acierto que cabe esperar cuando apliquemos el método a la identificación de peces cuya especie es desconocida.


Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana Ayuda
1 : Especie
1,00
Visible: 28 de 28 variab

	Especie	M2	M3	M4	M5	M6
64	Haemulon plumierii	,16	,05	,12	,08	,40
65	Haemulon plumierii	,15	,04	,13	,08	,36
66	Haemulon plumierii	,16	,05	,13	,09	,43
67	Haemulon plumierii	,16	,05	,12	,08	,39
68	Haemulon plumierii	,17	,05	,13	,10	,43
69	Haemulon plumierii	,18	,06	,16	,08	,40
70	Haemulon plumierii	,17	,05	,14	,08	,41
71	Haemulon plumierii	,15	,05	,12	,10	,41
72	Haemulon plumierii	,17	,07	,15	,07	,39
73	Haemulon steindachneri	,14	,04	,11	,10	,43
74	Haemulon steindachneri	,13	,04	,11	,10	,41
75	Haemulon steindachneri	,13	,03	,12	,11	,39
76	Haemulon steindachneri	,13	,03	,11	,10	,40
77	Haemulon steindachneri	,14	,04	,11	,10	,38
78	.	,13	,05	,11	,11	,41
79						

En la ventana del paso 2 pulsamos «Guardar» y en la ventana que aparece marcamos la opción «Grupo de pertenencia pronosticado». Se observa que aparecerá una nueva columna en la matriz de datos, donde el caso 78 se predice que pertenece a la especie *Haemulon melanurum*.



Capítulo XV

MODELOS DE SIMULACIÓN

XV.1. El uso de modelos

Los modelos se crean por la necesidad de expresar en lenguaje matemático cualquier tipo de razonamiento con el cual pretendamos analizar las propiedades de un sistema.

El lenguaje matemático es preciso y sus elementos son manipulables de manera que un modelo nos permite analizar particularmente todas las propiedades y componentes del sistema, y a la vez observarlo también como un todo.

Los modelos deben ser capaces de predecir con solvencia en ciertos niveles de generalidad, lo que significa que no deben ser útiles solamente para el caso concreto para el que fueron desarrollados.

Dentro de los muchos tipos de sistemas que han sido estudiados a partir de modelos, podemos encontrarnos desde sistemas físicos y químicos hasta económicos y biológicos. Se trata de campos de la ciencia muy diferentes. El modelado de sistemas de ciencias como la física o la química clásicas se ha desarrollado habitualmente desde perspectivas determinísticas, es decir, aquellas en las que los fenómenos son totalmente predecibles. Por el contrario, sistemas del tipo de los estudiados en economía o biología pueden moverse entre el determinismo y la estocástica (fenómenos de azar que dependen de una probabilidad) dependiendo siempre de la escala a la que estemos estudiando el sistema. En este segundo tipo de disciplinas, el modelado se ha incorporado de una manera más tardía, por el deseo de alcanzar un grado de precisión en estas ciencias similar al que existía en la física o la química, algo que hasta hoy ha sido difícil de conseguir debido a la complejidad de las relaciones y al carácter estocástico de algunos fenómenos que conforman los sistemas estudiados en estas disciplinas. Tenemos ejemplos de antiguos ensayos que estudiaban la dinámica de poblaciones humanas, como el de Malthus, desde una perspectiva de sus consecuencias económicas, pasando por los estudios fundacionales de modelado de sistemas ecológicos (Lotka, 1925, 1927; Volterra, 1926) hasta la actualidad, en la que todo tipo de sistemas han sido exhaustivamente estudiados (Ruth & Hannon 1997; Odum & Odum, 2000).

A la hora de modelar, es necesario un equilibrio entre simplicidad y predicción. Esto significa que un modelo debe ser simple de manejar matemáticamente, y a la vez reflejar con bastante aproximación qué cosas ocurren realmente en la naturaleza. Pero como estas dos premisas pueden ser incompatibles, es necesario encontrar una solución de compromiso.

Se ha aplicado el término "realista" (Gillman & Hails, 1997) en el sentido de la capacidad de un modelo para predecir el comportamiento de un sistema en la naturaleza. De forma antitética, el término "estratégico" (Gillman & Hails, 1997) se ha aplicado a modelos extremadamente sencillos que pretenden describir una pauta concreta en el funcionamiento de un sistema muy simple, pero que resulta irreal a la hora de explicar sistemas naturales.

Un ejemplo de modelo realista sería aquel en el que tenemos una serie de variables independientes que sabemos, por observación o experimentación, que afectan a la variable dependiente que queremos predecir. Si realizamos un modelo incluyendo todas estas variables y conseguimos una predicción razonablemente ajustada a lo que ocurre en la naturaleza, tendremos un modelo "realista". No obstante, en este tipo de modelos, por norma general, no sabemos bien cuál es el mecanismo que hace exactamente que cada variable afecte a nuestra variable dependiente. Por ejemplo, podemos construir un modelo en el que obtenemos una buena predicción de las capturas de una determinada especie de pez en función de los ciclos de actividad solar. Pero las causas proximales fisiológicas o las causas ecológicas por las que la actividad solar afecta a esa especie de pez no las conocemos.

Los modelos "estratégicos" por el contrario, definen muy claramente una pauta determinada que regula un proceso. Con ellos sabemos exactamente, de manera proximal, a qué se debe la dinámica que estamos observando. Un ejemplo de este tipo de modelos es el "depredador – presa" de Lotka–Volterra. Pero como ya mencionamos, se trata de modelos tan sencillos que nunca se cumplen sus predicciones en el medio natural.

La conveniencia de uno u otro tipo de modelos siempre va a depender de los objetivos de nuestra investigación.

Sea cual sea el tipo de modelo realizado, siempre nos permitirá "contextualizar" el sistema que es objeto de estudio, es decir, conocer todos sus componentes e individualizarlos moviéndonos en una o varias escalas en particular. Esto quiere decir que cualquier modelo que hagamos es una visión parcial de sistemas a mayor o menor escala, y precisamente el mérito de cualquier buen modelo consiste en saber acotar esa parte o subsistema que nos interesa estudiar, con todos sus componentes y en una escala o escalas adecuadas.

XV.2. Pasos a considerar para el desarrollo de un modelo

Brevemente, de manera muy generalizada, vamos a ordenar una serie de pasos que se deben dar en el proceso de desarrollo de un modelo.

1. Contextualización. En el origen de un modelo se encuentra la idea acerca del funcionamiento de un sistema que se crea en la mente del investigador.

    - Debemos establecer cuáles son los componentes de este sistema y cómo se relacionan.
    - Es necesario tener en cuenta la escala en la que actúa cada componente, un modelo puede hacerse todo en una escala o contener varias dentro de sí.
    - Por último dentro de este paso, debemos considerar siempre que nuestro sistema de interés es un pequeño aspecto que forma parte de otros sistemas más complejos, de manera que nuestros elementos siempre estarán conectados con procesos paralelos o en diferentes escalas.

2. Conformación de un diagrama. Esto se dirige ya al momento en que debemos emplear un determinado programa informático. Se trata de repre-

sentar gráficamente nuestro sistema obedeciendo a un lenguaje concreto de símbolos y unas leyes generales.

- Primero debemos realizar una lista de todos los elementos que contendría nuestro modelo. Por estos entendemos las variables, los flujos de energía, de fuerza y las conexiones ente ellos. Para esto, cada programa informático utiliza un tipo de lenguaje con diferentes símbolos, en el que cada uno representa un elemento con unas características definidas.
- Ahora procederíamos al diseño de los diferentes procesos que actúan en nuestro sistema, relacionando elementos.

3. Establecimiento de unidades, ecuaciones y parámetros. Cada proceso está controlado por una determinada ecuación. Cada ecuación se compone de una fórmula concreta que la caracteriza y de unos parámetros y variales. Todo esto se expresa en unas dimensiones o unidades concretas.

   A la hora de elegir las ecuaciones que controlan nuestros procesos, podríamos recordar algunas consideraciones:

   - Los procesos que estudiamos pueden ser determinísticos (predecibles) o estocásticos (dependientes de una probabilidad). A su vez, un fenómeno puede ser estocástico o determinístico en función de la escala a la que lo estudiemos.
   - Las unidades de tiempo pueden ser discretas o usarse tiempo continuo.
   - Si nuestro modelo contiene diferentes subprocesos en distintas escalas, debemos tener en cuenta que el aumento en escala conlleva un gran aumento en el número de los procesos.

   Para estimar el valor de los parámetros (tasa de crecimiento, mortalidad, capacidad de carga de una población, etc.) que introducimos en las diferentes ecuaciones del modelo, necesitaremos datos de situaciones reales. De lo bien medidos que estén estos datos dependerá en gran medida el éxito de nuestro modelo. En algunos casos puede no ser necesario conocer previamente un parámetro, por ejemplo, si lo que quieres buscar con el modelo es precisamente ese parámetro, conociendo el resto de los elementos.

   Las diferentes ecuaciones que introduzcamos, ya que están interconectadas, deben ser dimensionalmente homogéneas. Algunos programas informáticos, como el que usaremos nosotros, pueden comprobar si existen errores de este tipo.

4. Verificación del modelo. Es necesario, pero en muchos casos imposible por la falta de datos. A veces podemos suplir esta carencia por medio de ensayos experimentales.

XV.3. Introducción al modelado con el programa Stella

Una vez que tenemos claros los pasos a considerar para realizar un modelo, antes de proponer un ejemplo sobre el desarrollo de un modelo, necesitaremos una pequeña introducción para el manejo de un programa adecuado.

Hoy en día, el aumento de capacidad de los sistemas de computación ha facilitado en gran manera el desarrollo y la manipulación de modelos de gran complejidad, aumentando mucho su accesibilidad a la comunidad científica.

Existen diferentes tipos de programas creados específicamente para modelización dinámica (EXTEND, Imagine that; Stella, isee Systems), también es posible modelar con lenguajes de programación (BASIC, Maxima, S-plus, R, etc.) o bien en hojas de cálculo de diversos programas (EXCEL).

Nosotros vamos a explicar el manejo del programa Stella, por ser en la actualidad uno de los más adecuados para modelización dinámica que no requiere conocimientos de programación. Este programa permite a usuarios con niveles básicos de informática desarrollar modelos de bastante complejidad de una manera sencilla.

Los diferentes componentes (flujos, variables, constantes, etc.) se forman con distintos iconos que se conectan entre sí y el modelo se representa con un diagrama en la hoja de trabajo. Las ecuaciones que controlan los distintos procesos se escriben en dicho diagrama en el lugar que les corresponda.

Este programa tiene cuatro «capas» o «niveles»: *Interface*, *Map, Model y Equation*. *Map* y Model son los niveles básicos en que se diseña un modelo, creando un diagrama esquemático e introduciendo las ecuaciones que controlan los procesos representados en el diagrama. El *Interface* es un nivel para usuarios avanzados en el que un modelo creado en el *Map/Model* se puede hacer interaccionar, a medida que la simulación se produce, con efectos controlados por el usuario. *Equation* simplemente es un apartado donde se describen detalladamente todos los componentes del modelo que hemos creado.

Haremos una explicación a nivel de usuario básico a través de un modelo sencillo, que será necesaria para después entender el lenguaje y funcionamiento del modelo ecológico que propondremos en el apartado XV.4.

XV.3.1. SIGNIFICADO DE LOS ICONOS Y MENÚS ESPECÍFICOS DE LA BARRA DE HERRAMIENTAS EN NIVELES *MODEL, MAP* E *INTERFACE*

En primer lugar explicaremos el significado de los iconos del programa en los niveles *Model y Map.*

Stock: depósitos en los que otras variables modifican el contenido. Hay varios tipos, que se seleccionan al mantener pulsado el icono, y son los siguientes:

1. *Conveyor*: funciona como una cinta transportadora, en la que los elementos permanecen un tiempo y salen.
2. *Queue*: funciona como una cola, en la que los elementos van entrando y saliendo por el mismo orden.
3. *Oven*: los elementos entran y salen en cantidades discretas; los de cada grupo entran y salen a la vez.
4. *Submodel*: submodelo que funciona como stock dentro de otro.

Flow: crea flujos de entrada o de salida sobre un *Stock*. El standard es unidireccional, pero manteniendo el icono pulsado tenemos la opción de crear uno bidireccional.

Converter: valor o expresión cuyos efectos modifican otros elementos del modelo como *Flows* o *Stocks*. Manteniendo pulsado el icono podemos crear un *summer converter* (ver cuadro XV.1, Paso 9).

Action connector: establece relaciones entre varios elementos. Manteniendo el icono pulsado permite seleccionar un Action connector de información.

Module: un módulo sirve para conectar modelos jerárquicamente. Así, un modelo puede dividirse en "submodelos" que se construyen dentro de un módulo, y que se conectan entre sí. Si mantenemos este icono pulsado, aparece debajo otro denominado "Decision Proccess Diamond" que sirve para establecer reglas para la toma de decisiones.

Button: diseñado para simplificar el modo de realizar diversas tareas dentro de un modelo. Un botón puede crearse con el objetivo de que, al pulsarlo, se ejecute una orden por comandos, una navegación, o el despliegue de una determinada información, etc.

Sector Frame: en modelos complejos, permite agrupar en un sólo sector elementos relacionados funcionalmente, simplificando las conexiones del modelo.

Graph pad: permite visualizar en un gráfico los resultados de una simulación.

Table pad: permite visualizar en una tabla los resultados de una simulación.

Status indicator: nos permite deplegar información acerca del estado de un determinado "output" de nuestro modelo durante la simulación.

Numeric display: permite mostrar, en una pequeña ventana, los valores que toma un elemento durante la simulación.

Text: se usa para escribir un texto.

Graphic frame: podemos situar en él gráficos o películas importadas de otros archivos.

Arrow: selecciona iconos para operar con ellos en la barra y la hoja de trabajo y otras funciones.

Paintbrush: sirve para colorear elementos del modelo.

Dynamite: elimina cualquier elemento seleccionado en la hoja de trabajo.

Ghost: sirve para situar una copia igual de un elemento del modelo en otro lugar alejado del original, y así facilitar gráficamente su conexión con otros elementos.

Dentro de *Interface*, disponemos de algunos iconos diferentes.

Proccess frame: Facilita la representación de procesos a nivel superior en el modelo. Especialmente cuando se trata de modelos estructurados con un tipo de control predominante "de arriba hacia abajo", es decir, de los niveles superiores a los inferiores.

Bundled flow: permite el flujo entre niveles superiores de un modelo.

Bundled connector: representa a modo de resumen todas las conexiones existentes entre diferentes sectores del modelo.

Loop pad: permite representar en un diagrama las estructuras causales que constituyen un bucle existente dentro del modelo.

Switch: este "interruptor" puede asignarse a diferentes componentes del modelo para, en un momento determinado, activarlos o apagarlos.

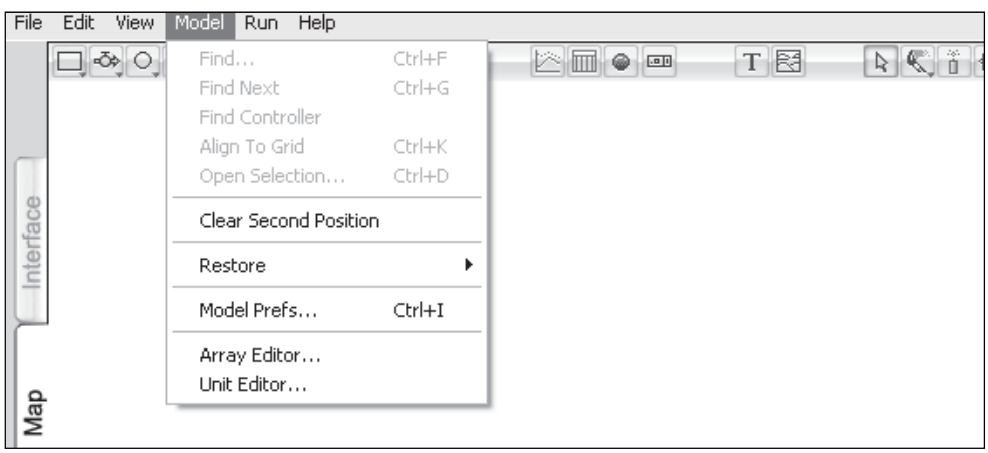
Knob input device: en general se usa para determinar valores iniciales de stocks. También se puede usar para dar valores a constantes.

List input device: en un formato tipo hoja de cálculo, permite introducir valores para *converters*, *flows* y valores iniciales de *stocks*. Una sola de estas listas puede contener valores y datos para varios componentes del modelo a la vez.

Slider input device: sirve para ajustar los valores de las constantes sin tener en cuenta la lógica de las ecuaciones ni las funciones gráficas cuando así nos interese.

Ahora haremos un resumen del significado de los menús más importantes del programa.

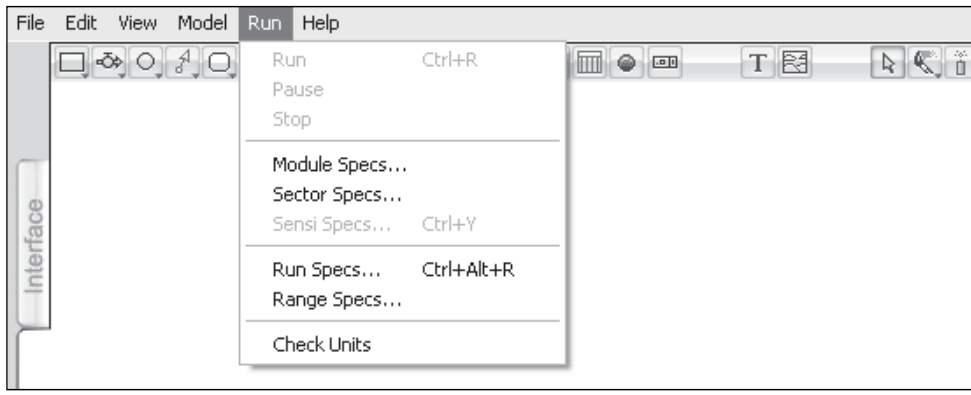
Model**:** Sus comandos se refieren a aspectos de formato, como buscar elementos concretos del modelo, seleccionarlos, esconderlos o volver a mostrarlos y juntar gráficamente en uno sólo varios procesos paralelos (comando *Array Editor*), por ejemplo, si una empresa de aserraderos tiene 5 naves donde trabajan, el proceso que regula la entrada de madera es el mismo y es sencillo, pero gráficamente puede estorbar ver 5 *Stocks* paralelos, entonces, con *Array Editor*, los condensamos en uno con efectos en 3 dimensiones.



Run**:** Configuración y puesta en marcha de una simulación. Hay que destacar el comando "*Run Specs…*", en donde se controlan las unidades de tiempo, tipo de integración y si queremos incluir interacciones (sólo para el *Flight Simulator*, donde las interacciones se crean sobre el modelo a medida que se simula). Es importante aquí el método de integración, para el que existen tres variantes:

1. *Euler*: funciona por defecto, es el adecuado cuando en el modelo hay objetos "discretos", tales como funciones lógicas (IF, ELSE, etc.) o ROUND, SWITCH y otros.
2. *Runge–Kutta* 2 y 4: son los más adecuados para procesos de cambio contínuo, con tendencia inherente a la oscilación y/o ciclos, en los que el método de Euler produce errores acumulativos ciclo tras ciclo.

El menú "*Run*" también permite simular o configurar sólo un sector del modelo y asignar rangos a las variables.

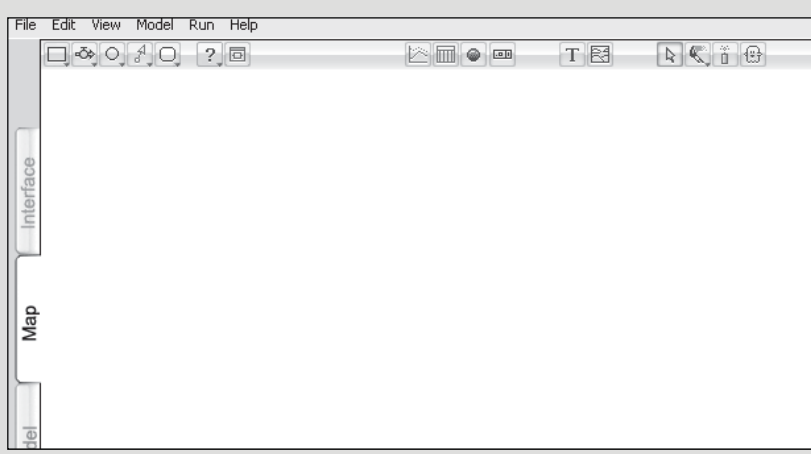


XV.3.2. EJEMPLO DE FUNCIONAMIENTO DE LOS ICONOS Y MENÚS BÁSICOS

Vamos a modelar la explotación sostenible de un acuífero, en el que no queremos reducir sus reservas de agua por debajo de una cantidad que consideramos adecuada. Las ecuaciones y cálculos matemáticos se simplificarán extremadamente. Podemos verlo en el Cuadro XV.1.

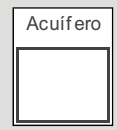
CUADRO XV.1. Modelo de explotación sostenible de un acuífero

Paso 1. Tenemos abierta la pantalla de Stella, que por defecto se abre en *Map.*

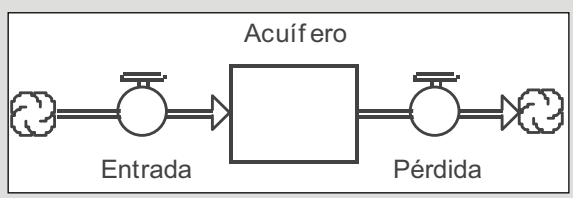


CUADRO XV.1. *(Continuación)*

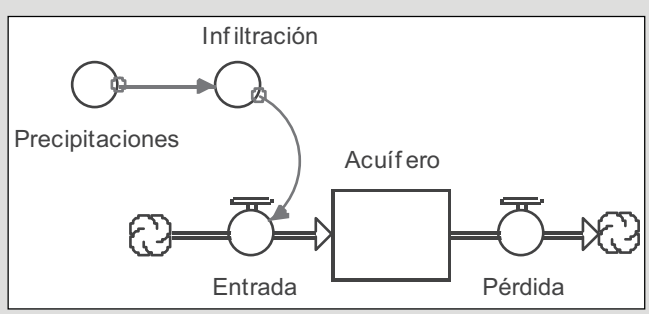
Paso 2. Pulsamos sobre el icono de *Stock* y luego sobre la hoja de trabajo, así lo situamos en ésta. Le cambiamos el nombre de *Noname1* haciendo doble clic en la etiqueta, y lo llamamos «Acuífero».



Paso 3. El acuífero tiene flujos de entrada de agua y flujos de pérdida. Para representarlos pulsamos sobre el icono *Flow* y luego en la hoja de trabajo. *Flow* es una flecha, para situarla se hace clic en la hoja de trabajo y no se suelta el ratón hasta extender la flecha hacia donde queremos. El flujo de alimentación lo hacemos situando el origen de la flecha fuera del *Stock* y su extremo conectado con él. El flujo de pérdida se hace al revés, con el origen conectado al *Stock* y el extremo hacia fuera de él. Nombramos a los flujos «Entrada» y «Pérdida».

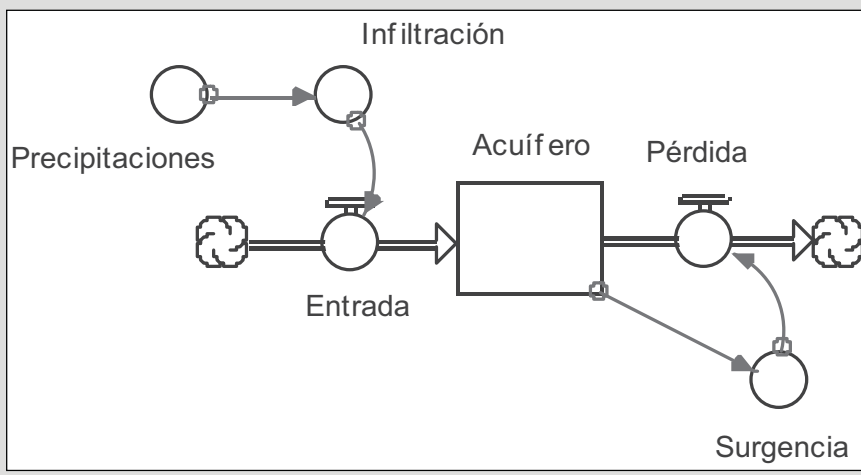


Paso 4. Las entradas de agua en el acuífero provienen de la infiltración hasta la capa freática. Representaremos la infiltración como un *Converter*, que uniremos con un *Action connector* al flujo de entrada. A su vez, las precipitaciones influyen en la infiltración, por lo que las representaremos como un *Converter* unido a la infiltración por otro *Action connector*. Los *Action connectors* se extienden desde el icono de origen hasta el de destino. Nombramos a cada elemento.

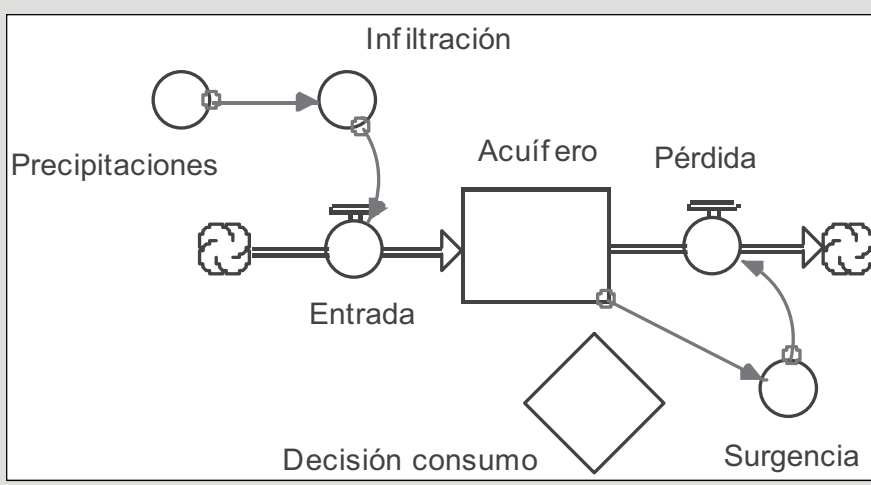


CUADRO XV.1. *(Continuación)*

Paso 5. Las pérdidas por surgencia natural del acuífero afectarán al flujo de pérdidas. Esta surgencia natural de agua sería un *Converter* unido al flujo de pérdida por un *Action connector*, y también al Stock «Acuífero», pero con el *Action connector* al revés que antes, ya que el caudal de surgencia depende del nivel de agua del acuífero.

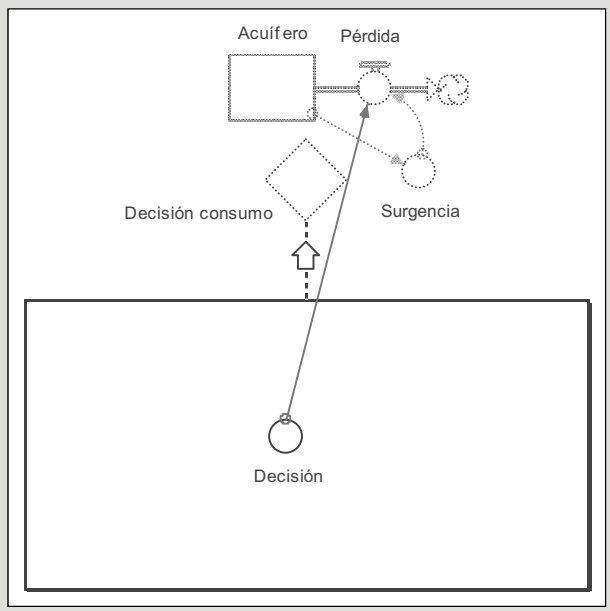


Paso 6. El consumo de agua afecta a las pérdidas, pero queremos controlarlo, por ello debemos introducirlo como un proceso controlable en función de la entrada y las pérdidas naturales. Para esto, el icono con una función más indicada es el *Decision Process Diamond* (*DPD*) (ver *Module* en el texto), lo nombramos como «Decisión consumo».

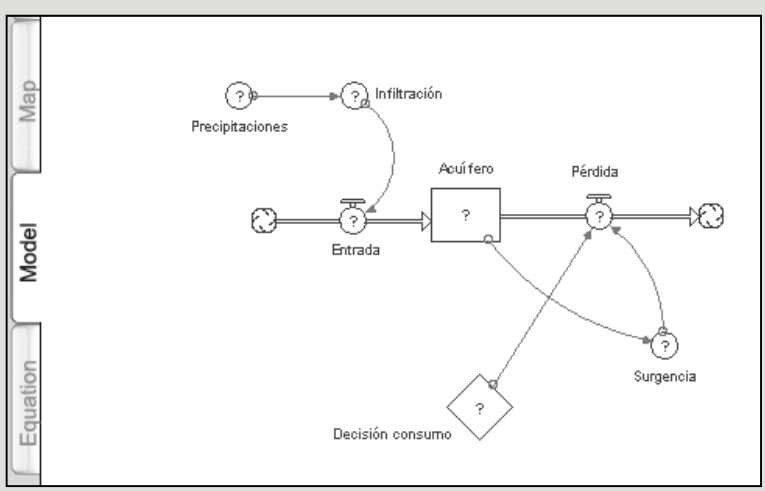


CUADRO XV.1. *(Continuación)*

Paso 7. Ahora pulsamos dos veces dentro del icono del *DPD*, y se abre un área rectangular (ver diagrama abajo). En esta área podríamos situar cualquier mini-diagrama o expresión de la cuál obtengamos una decisión. Para hacerlo, tenemos disponible la información de las variables en línea discontinua. Situamos un *Converter* al que llamaremos Decisión, y lo unimos al flujo de pérdida.

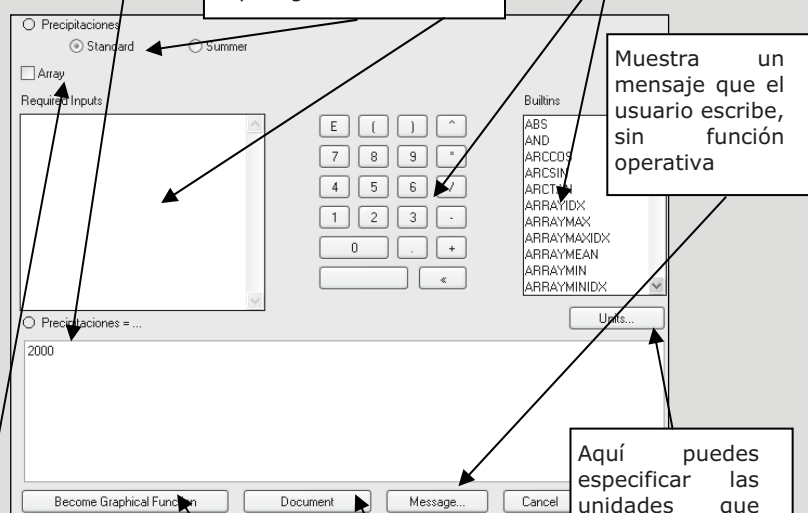


Paso 8. Pulsamos en la pestaña *Model*. Aquí ya podemos introducir las funciones matemáticas. Aparecerá una interrogación sobre cada elemento hasta que introduzcamos en él una ecuación, constante o expresión lógica.



CUADRO XV.1. *(Continuación)*

Paso 9. Introducimos primero el valor de las precipitaciones. Para ello hacemos doble clic en el icono del diagrama con ese nombre y aparece el siguiente cuadro de diálogo, en donde introduciríamos la expresión matemática que le corresponde.



Espacio para introducir la expresión. En nuestro caso una constante de dimensiones litros/unidad de tiempo y con valor 2000

Standard nos muestra en *Required inputs* los elementos del modelo conectados con este *Converter* que se deben introducir en la función. *Summer*, nos permitiría introducir también los no conectados, pero sólo aparece cuando no hay ninguno conectado

Con estos operadores y funciones podemos construir la expresión matemática que queramos introducir

Muestra un mensaje que el usuario escribe, sin función operativa

Con Array simplificamos en un solo icono varios procesos paralelos

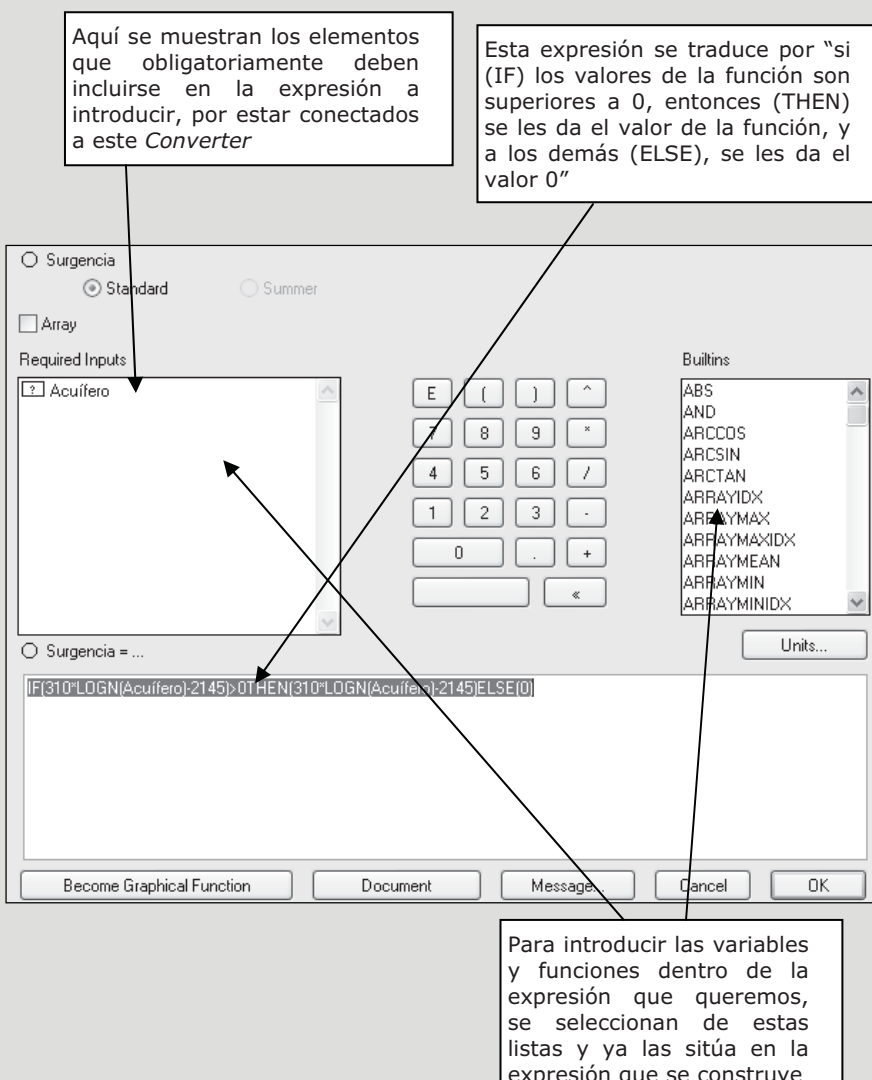
Aquí puedes especificar las unidades que utilizas y comprobar la homogeneidad dimensional de la expresión que has introducido. No lo usaremos

Aquí se pueden transformar los valores que calcula la expresión que has introducido a través de una función gráfica, que definiríamos en un cuadro de diálogo que se abre pulsando aquí. No lo usaremos

Nos permitiría acceder a un *Document*. Este se crea, si es necesario, en *mapping mode* haciendo doble clic sobre la variable y se escriben las asunciones que corresponden a la variable

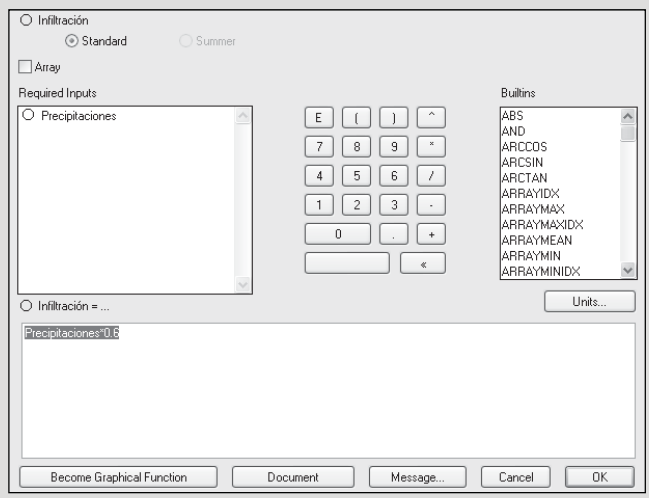
CUADRO XV.1. *(Continuación)*

Paso 10. Ahora abrimos el cuadro de «Surgencia». Entonces introducimos una expresión calculada experimentalmente que es la función que relaciona el valor de esta variable con las reservas del acuífero. Esta función sería «310*LOGN(Acuífero)-2145». Pero como esta función se hace negativa en valores bajos, tenemos que especificar al modelo que en lugar de estos valores negativos, le asigne a la función un valor de 0. Para esto introducimos una combinación de expresiones lógicas y la función anterior, como podemos ver.

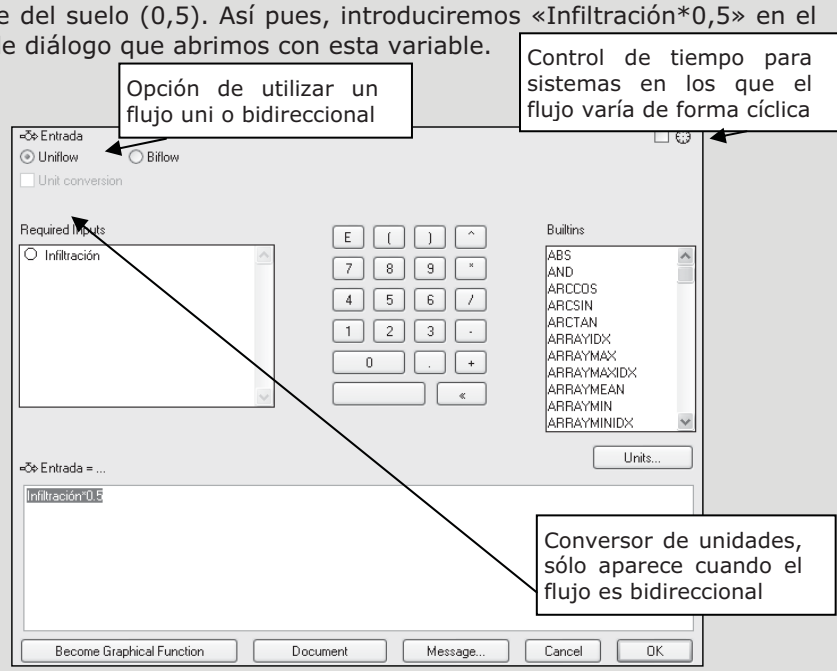

Aquí se muestran los elementos que obligatoriamente deben incluirse en la expresión a introducir, por estar conectados a este *Converter*
Esta expresión se traduce por "si (IF) los valores de la función son superiores a 0, entonces (THEN) se les da el valor de la función, y a los demás (ELSE), se les da el valor 0"
Surgencia
Standard
Summer
Array
Required Inputs
Acuífero
Builtins
ABS
AND
ARCCOS
ARCSIN
ARCTAN
ARRAYIDX
ARRAYMAX
ARRAYMAXIDX
ARRAYMEAN
ARRAYMIN
ARRAYMINIDX
Units...
Surgencia = ...
IF(310*LOGN(Acuífero)-2145)>0THEN(310*LOGN(Acuífero)-2145)ELSE(0)
Become Graphical Function
Document
Message...
Cancel
OK
Para introducir las variables y funciones dentro de la expresión que queremos, se seleccionan de estas listas y ya las sitúa en la expresión que se construye

CUADRO XV.1. *(Continuación)*

Paso 11. Sabemos que el volumen de infiltración de agua es una función lineal que depende de las precipitaciones y de un coeficiente de permeabilidad del suelo, que en nuestro caso es 0,6. Entonces, hacemos doble clic en «Infiltración», se abre el cuadro de diálogo siguiente e introduciríamos la expresión: «Precipitaciones*0,6» como se observa abajo.

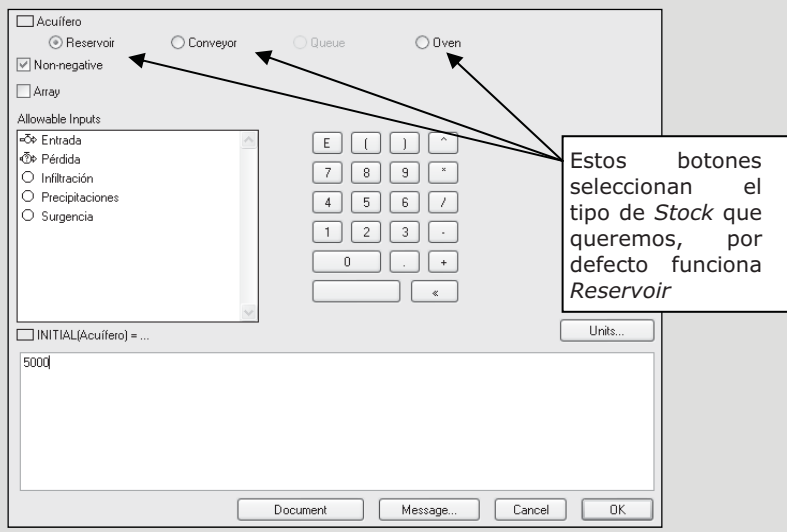


Paso 12. Haremos lo mismo con la variable «Entrada», que viene dada por una relación lineal entre la infiltración y un coeficiente de retención de agua por parte del suelo (0,5). Así pues, introduciremos «Infiltración*0,5» en el cuadro de diálogo que abrimos con esta variable.

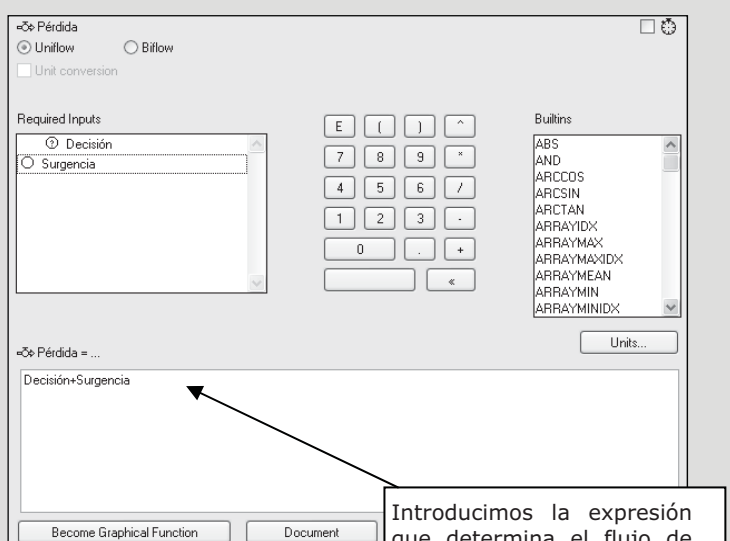


CUADRO XV.1. *(Continuación)*

Paso 13. Ahora introduciremos el valor del *Stock*, es decir, los litros de agua que contiene el acuífero a medida que las entradas y pérdidas le afectan. Hacemos doble clic en el *Stock* y escribimos 5000 litros en el cuadro de diálogo.

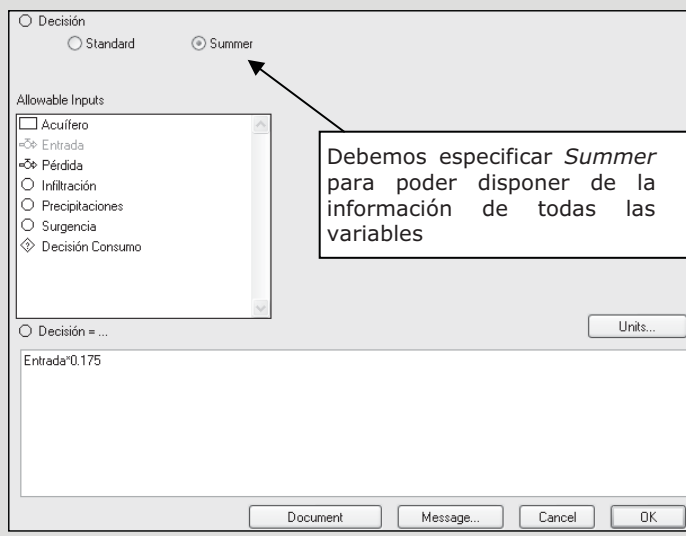


Paso 14. También introducimos ahora el flujo de «Pérdidas» que vendrá dado por la suma del agua perdida por surgencia y por consumo. Hacemos doble clic en la variable y configuramos el cuadro de diálogo como se ve en la figura. El cuadro es idéntico al de «Entrada».

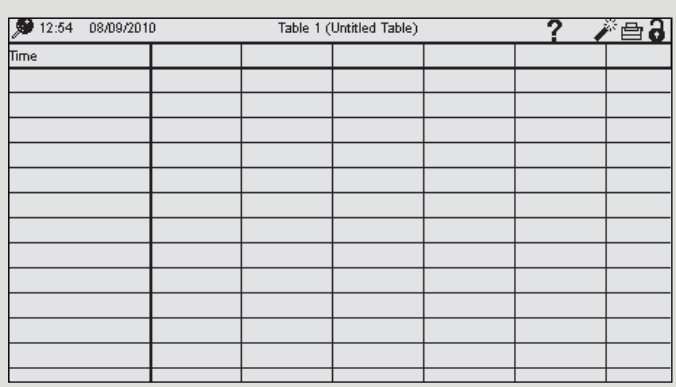


CUADRO XV.1. *(Continuación)*

Paso 15. Ahora ya podemos introducir la expresión que nos indicará cuánta agua se puede consumir. Abrimos el *DPD*, y sobre el área rectangular hacemos doble clic sobre «Decisión» y se abre el cuadro que mostramos abajo. El consumo debe tener en cuenta las entradas y las pérdidas, manteniendo equilibrados ambos valores. Además, queremos mantener el acuífero con un valor de surgencia de aproximadamente 500 (que se alcanza con 5000 litros en el acuífero) porque estimamos que ese sería el caudal adecuado para no alterar los sistemas naturales que dependen de este caudal de agua. Calculamos que, para esto, la proporción de agua consumida debe ser la fracción 0,175 de las entradas. Así pues, la expresión que introduciremos es «Entrada*0,175».

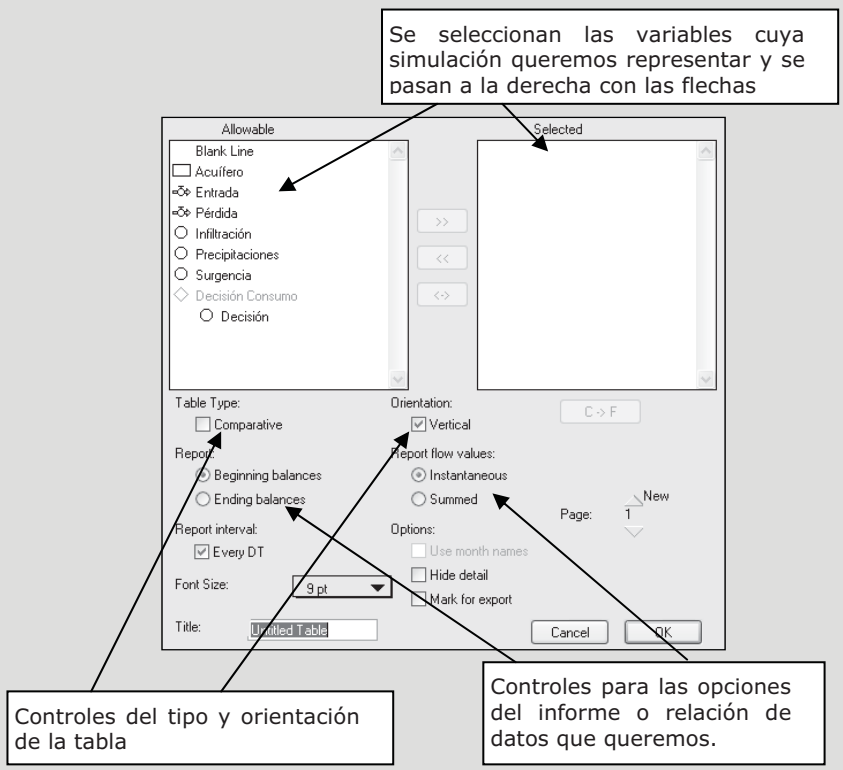


Paso 16. Ahora cerramos el *DPD* haciendo clic en la flecha y realizaremos una simulación. No pueden visualizarse a la vez la tabla y el gráfico, así que primero lo veremos en la tabla. Para esto hacemos clic en el icono *Table Pad*, y lo situamos en la hoja de trabajo, hacemos doble clic y abrimos la tabla que aparece a continuación.

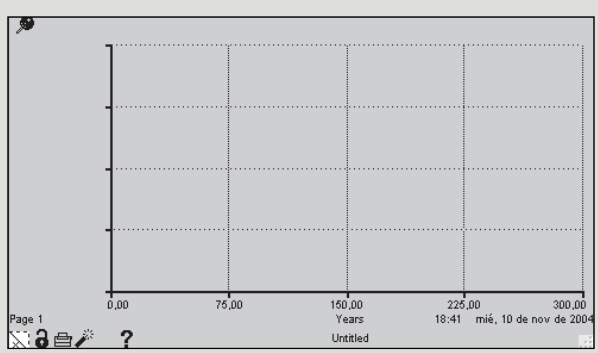


CUADRO XV.1. *(Continuación)*

Paso 17. Hacemos doble click sobre la primera columna y aparece el siguiente cuadro de diálogo, que configuraremos como se indica, para obtener en la simulación los datos que corresponden a cómo varía el contenido del acuífero y la surgencia.

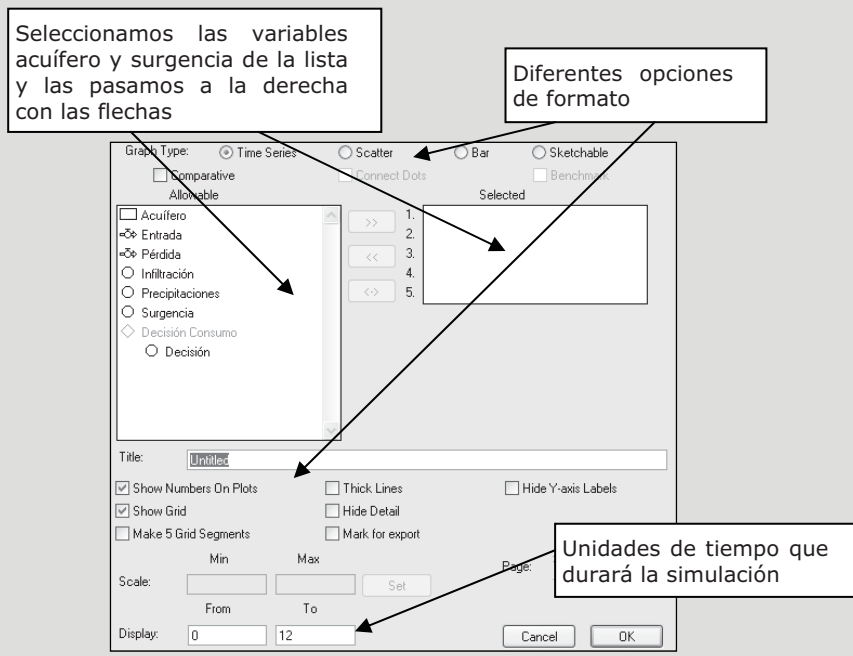


Paso 18. Para observar la simulación gráficamente, hacemos clic en el icono *Graph Pad* y colocamos el gráfico siguiente en la hoja de trabajo.

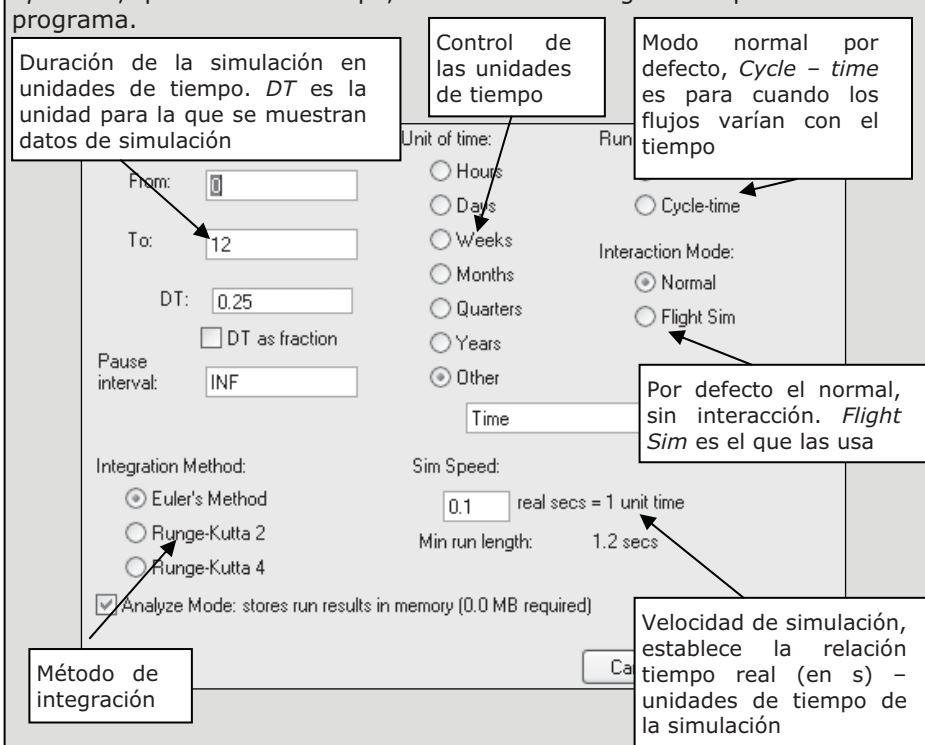


Paso 19. Para configurarlo, hacemos doble clic en el área del gráfico y nos abre el siguiente cuadro de diálogo.

CUADRO XV.1. *(Continuación)*



Paso 20. Ahora pulsamos sobre el menú «*Run*». En el comando «*Run Specs…*», que mostramos aquí, usaremos la configuración por defecto del programa.

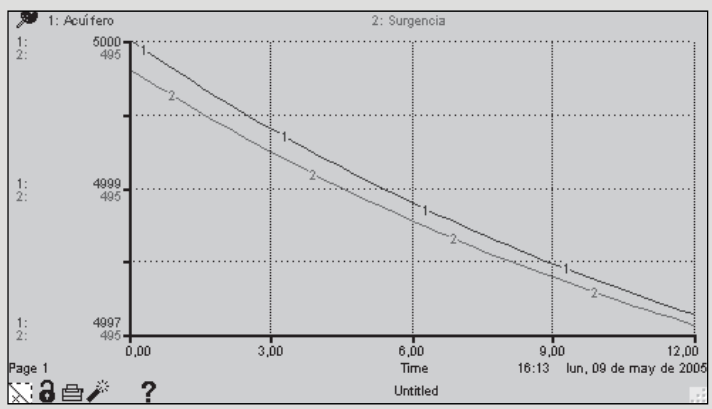


CUADRO XV.1. *(Continuación)*

Paso 21. Para comenzar la simulación pulsamos sobre el comando *Run* del menú «*Run*». Como dijimos antes, no se pueden abrir a la vez la tabla y el gráfico. Abriéndolos por separado y haciendo la simulación de cada vez, observaríamos lo siguiente.

13:14 08/09/2010 Table 1 (Untitled Table)

Time	Acuífero	Surgencia				
,00	5.000,00	495,33				
,25	4.999,92	495,32				
,50	4.999,84	495,32				
,75	4.999,76	495,31				
1,00	4.999,68	495,31				
1,25	4.999,60	495,31				
1,50	4.999,52	495,30				
1,75	4.999,45	495,30				
2,00	4.999,37	495,29				
2,25	4.999,30	495,29				
2,50	4.999,23	495,28				
2,75	4.999,16	495,28				
3,00	4.999,09	495,27				



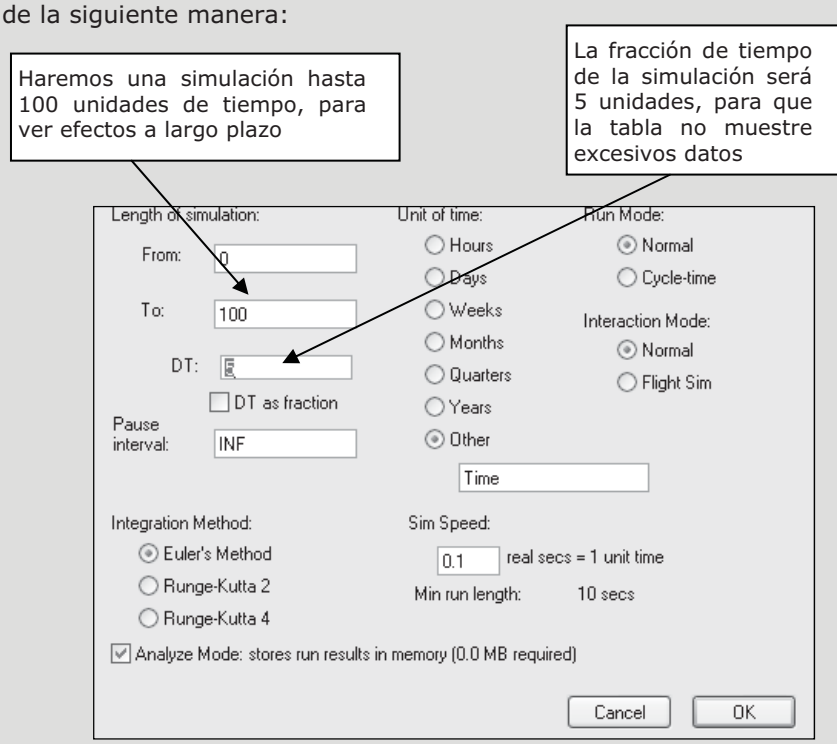
Las reservas del acuífero y la surgencia se mantendrían casi constantes, consumiendo aproximadamente un 17% del agua que es renovada. Podemos ver el modelo en el archivo **Modelo final STM**.

Con este tipo de modelos y en este programa cabe la posibilidad, de una manera muy sencilla, de predecir respuestas del sistema ante cambios en las variables por condiciones ambientales, y así poder adaptar nuestras hipotéticas actuaciones. Por ejemplo, ¿qué sucedería si en la unidad de tiempo considerada llueve un 50% menos? o ¿qué sucedería si la permeabilidad del suelo se reduce por impactos antropogénicos y la tasa de infiltración cae un 30%? Se trataría de modificar los valores o expresiones de estas variables y simular. Vamos a ver qué sucedería en el primero de los ejemplos propuestos.

Paso 22. Para esto, en primer lugar debemos introducir 1000 litros en «Precipitaciones», de la misma forma que antes 2000.

CUADRO XV.1. *(Continuación)*

Paso 23. Ahora realizaremos la simulación, pero antes, en el menú «*Run*», comando «*Run Specs…*», modificamos la configuración por defecto de la siguiente manera:

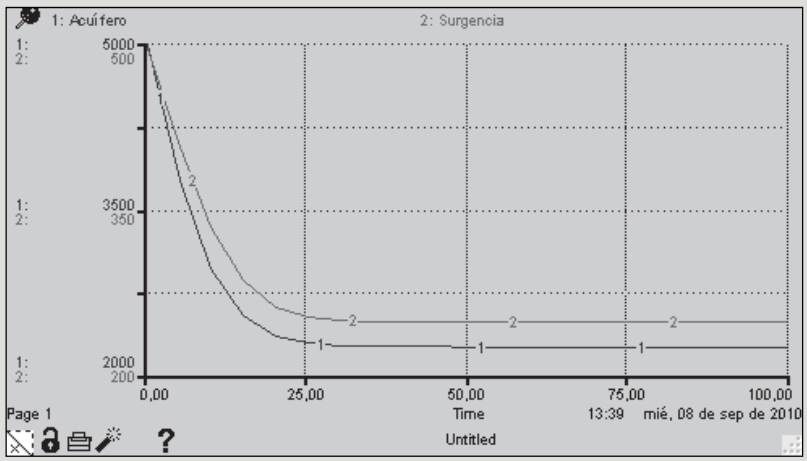


Paso 24. Una vez realizado esto, hacemos la simulación. Obtenemos los siguientes valores en la tabla y la gráfica.

13:23 08/09/2010 — Table 1 (Untitled Table)

Time	Acuífero	Surgencia						
0	5.000,00	495,33						
5	3.760,85	407,04						
10	2.963,13	333,14						
15	2.534,93	284,76						
20	2.348,65	261,09						
25	2.280,68	251,99						
30	2.258,22	248,92						
35	2.251,11	247,94						
40	2.248,88	247,64						
45	2.248,19	247,54						
50	2.247,97	247,51						
55	2.247,91	247,50						
60	2.247,89	247,50						
65	2.247,88	247,50						
70	2.247,88	247,50						
75	2.247,88	247,50						
80	2.247,88	247,50						
85	2.247,88	247,50						
90	2.247,88	247,50						
95	2.247,88	247,50						
Final	2.247,88	247,50						

CUADRO XV.1. *(Continuación)*



Como vemos, las reservas del acuífero disminuyen, pero no llegan a agotarse, sino que se hacen constantes en torno a 2200 litros, porque la surgencia, aunque disminuye mucho, no llega a agotarse, y el consumo, aunque es excesivo para mantener el caudal óptimo de 500 litros/unidad de tiempo de surgencia, no llega a secar el acuífero.

La pretensión de este ejemplo no es más que una introducción al manejo del programa desde el punto de vista del uso de sus funciones y desarrollo conceptual, siempre a nivel muy básico. El objetivo es aprender a manejarlo para poder desarrollar un modelo como el propuesto en el apartado XV.4.

XV.4. Ejemplo de modelado con el programa Stella: desarrollo conceptual y manejo práctico

Ahora que ya sabemos los pasos que hay que dar para realizar un modelo y manejar el programa Stella en sus aspectos fundamentales, vamos a proponer un ejemplo de cómo se desarrolla la creación de un modelo, al que llamaremos *modelo de la interacción depredador – presa*.

Este tipo de modelo fue diseñado por Lotka y Volterra, pero sabemos que en pocos casos se puede observar esta dinámica en condiciones naturales.

En el ejemplo iremos introduciendo paso a paso mayor complejidad en el modelo, en concreto, lo dividiremos en tres pasos. En el primer paso simplemente tendremos en cuenta la dinámica poblacional de una presa sin ninguna otra interacción. Después le añadiremos la dinámica poblacional del depredador y su influencia sobre la presa. Por último, añadiremos al modelo un efecto de explotación sobre las poblaciones del depredador, que aumentará su mortalidad.

Una de las ideas fundamentales que provocaron la búsqueda de un modelo de este tipo fue el considerar que muchos organismos que dependan de otros para

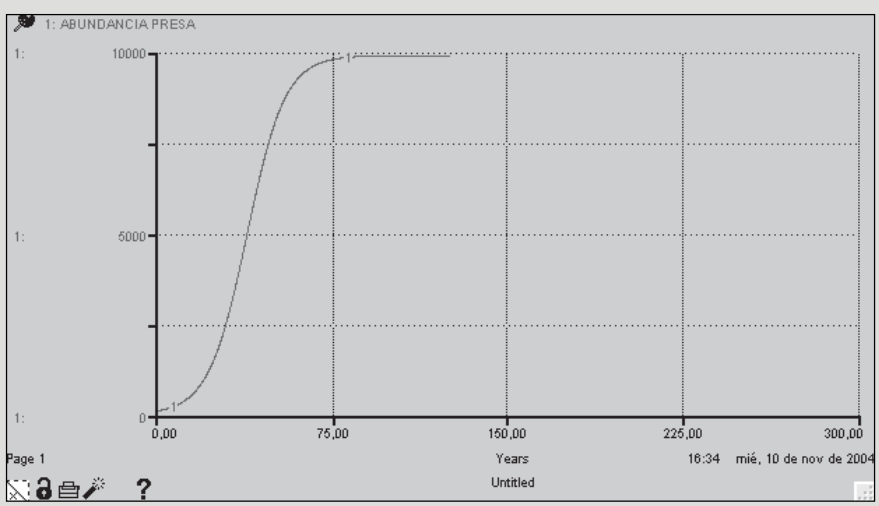
alimentarse (como por ejemplo depredadores especialistas y sus presas) deben producir fuertes impactos en sus poblaciones debidos al consumo. A su vez, las presas, en función de este y otros factores que afecten a su dinámica poblacional, provocarán un efecto sobre la población de los depredadores. Todo esto, intuitivamente, sugiere que los organismos que se relacionen de esta manera deben sufrir cambios cíclicos en sus poblaciones, y se trata de encontrar el modelo que los describa.

XV.4.1. DINÁMICA POBLACIONAL DE LA PRESA

En el Cuadro XV.2 podemos encontrar las especificaciones del modelo para este primer paso y el resultado gráfico. Ahora se trata de que, aplicando los conocimientos que hemos adquirido hasta ahora, cada uno sea capaz de crear un modelo con esas características.

CUADRO XV.2. Especificaciones del modelo dinámica poblacional de la presa y resultado gráfico

Método de Integración	Método de Euler
Unidad de tiempo	Meses
DT	0,1 meses
Tasa de crecimiento poblacional presa (r)	2
Capacidad de carga poblacional presa (K)	100000 individuos
Crecimiento poblacional presa	$\frac{N(K-N)}{K} \times e^{rt} - N$
Mortalidad presa	0,1 x N (individuos)
Abundancia inicial presa	100 individuos



Vamos a justificar la utilización de cada una de las ecuaciones. Nuestra primera consideración es que el tiempo es continuo, porque la reproducción de nuestra presa ocurre de esta manera, necesitaremos pues, una ecuación diferencial.

Por otro lado, el tipo de crecimiento de nuestra población será dependiente de la densidad, es decir, presentará un crecimiento más lento a bajas densidades, debido a la menor probabilidad de encuentro entre individuos, una fase intermedia en la que el crecimiento es máximo, y una fase final en la que la densidad se estabiliza en un valor concreto que es la capacidad de carga de nuestra población, y se debe al efecto de la competencia intraespecífica, que imposibilita que el medio soporte más individuos.

Nuestra especie de presa se considera que puede crecer sin limitación de alimento, ya que en el medio encuentra todo el que necesita, mientras no se llegue a los valores críticos de competencia intraespecífica.

La ecuación que utilizamos para definir esta dinámica es la ecuación de crecimiento logístico. Esta ecuación tiene dos parámetros que son la tasa de crecimiento *r* y la capacidad de carga *K*, que deberemos estimar a partir de medidas experimentales.

Otro parámetro que debemos incluir aquí es la mortalidad de la presa, que en nuestro caso consideraremos constante para simplificar, con lo que la ecuación es inmediata.

Podemos ver cómo quedaría este modelo en **Modelo 1d.STM**.

XV.4.2. DINÁMICA POBLACIONAL DEL DEPREDADOR Y SU INFLUENCIA SOBRE LA PRESA

Introducimos ahora un depredador que nunca sería capaz de sobrevivir en nuestro sistema si no tuviese a la presa. La idea es que, como decíamos antes, la dinámica poblacional de ambos se interrelacionará para obtener procesos cíclicos. En el Cuadro XV.3 detallamos las especificaciones del modelo en este nuevo paso.

La ecuación de crecimiento de la población del depredador será la misma que la de la presa, por idénticos motivos.

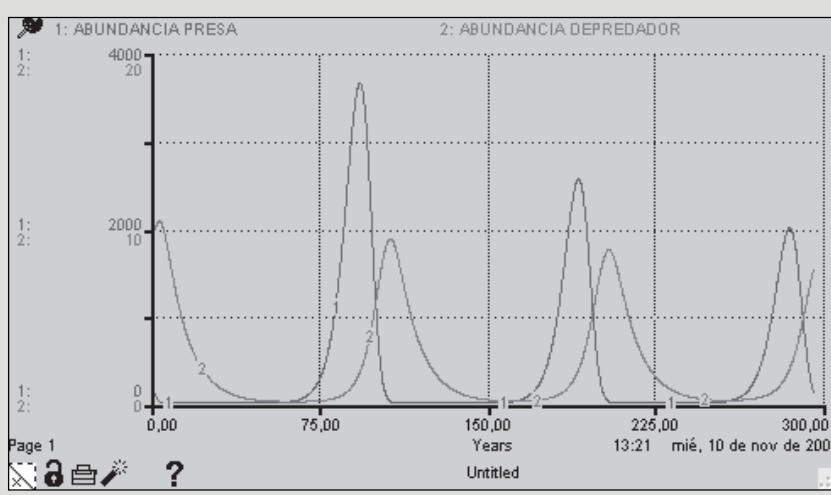
La mortalidad de la presa se asume que depende de una proporción 0,1*abundancia del depredador*abundancia de la presa, de manera que ya no es constante, como en el paso inicial sin depredador.

Esta mortalidad de la presa provocada por el consumo del depredador conlleva un aumento de la densidad del depredador, reflejado en el diagrama de Stella por las conexiones de la mortalidad de la presa con la abundancia de la presa y esta abundancia, con la tasa de crecimiento poblacional del depredador.

Así pues, la densidad del depredador crece por el consumo de la presa hasta llegar a un punto que hace aumentar la mortalidad de la presa hasta que esta supera al crecimiento de la misma. Esto provoca una disminución en la abundancia de la presa, y consecuentemente, una disminución retardada en la población del depredador.

CUADRO XV.3. Especificaciones del modelo dinámica poblacional del depredador y su influencia sobre la presa y resultado gráfico

Método de Integración	Método de Euler
Unidad de tiempo	Meses
DT	0,1 meses
Abundancia inicial presa	100 individuos
Tasa de crecimiento poblacional presa (r_p)	2
Capacidad de carga poblacional presa (K_p)	100000 individuos
Crecimiento poblacional presa	$\frac{N(K-N)}{K} \times e^{rt} - N$
Mortalidad presa	0,1 x *N* presa x *N depredador*
Abundancia inicial depredador	10 individuos
Tasa de crecimiento poblacional depredador (r_d)	0,3 ln (*N presa* +1)
Capacidad de carga poblacional depredador (K_d)	100000 individuos
Crecimiento poblacional depredador	$\frac{N(K-N)}{K} \times e^{rt} - N$
Mortalidad depredador	0,1 x *N* depredador



Llegaríamos a un punto en el que la densidad del depredador es tan baja, que la mortalidad de la presa vuelve a ser superada por su tasa de crecimiento, volviendo a aumentar la densidad de la presa, y con ella, un poco retardada, la del depredador.

Como vemos, todas las propiedades de la dinámica poblacional inicial han cambiado y hemos llegado a una situación en la que tiene lugar, si observamos la simulación, un proceso cíclico (ver Cuadro XV.3). El modelo puede verse en **Modelo 2d.STM**.

XV.4.3. INFLUENCIA DE LA EXPLOTACIÓN SOBRE LA POBLACIÓN DEL DEPREDADOR

Ahora introduciremos un efecto de explotación sobre la población del depredador. Podemos ver las nuevas especificaciones en el Cuadro XV.4.

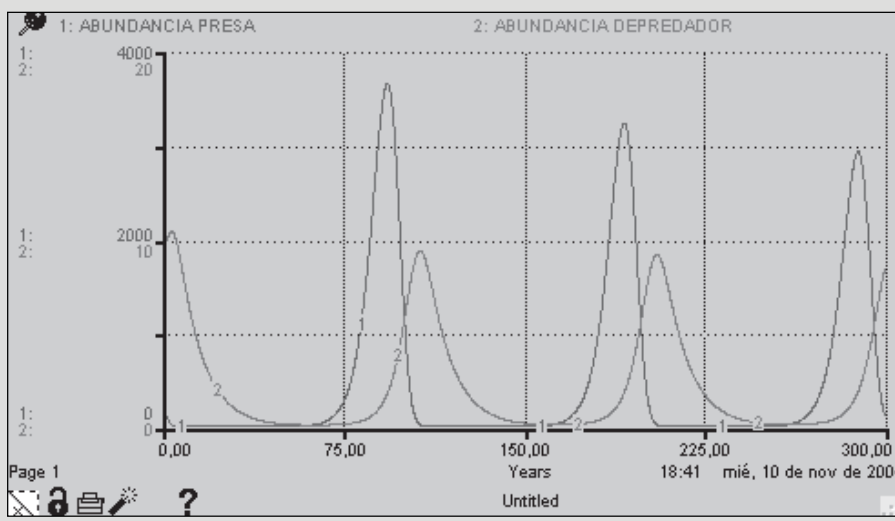
Este efecto de explotación aumentará la mortalidad del depredador. En nuestro caso asumimos que las capturas son constantes. Para ver juntas las dos fases del modelo (con y sin explotación) hemos incluido el efecto de explotación a partir del año 100. El único efecto que tiene esto es una mortalidad más rápida del depredador, que acorta la duración de los ciclos. Podemos verlo si hacemos la simulación en **Modelo 3d.STM**.

CUADRO XV.4. Especificaciones del modelo influencia de la explotación sobre la población del depredador y resultado gráfico

Método de Integración	Método de Euler
Unidad de tiempo	Meses
DT	0,1 meses
Abundancia inicial presa	100 individuos
Tasa de crecimiento poblacional presa (r_p)	2
Capacidad de carga poblacional presa (K_p)	100000 individuos
Crecimiento poblacional presa	$\frac{N(K-N)}{K} \times e^{rt} - N$
Mortalidad presa	0,1 x *N* presa x *N depredador*
Abundancia inicial depredador	10 individuos
Tasa de crecimiento poblacional depredador (r_d)	0,3 ln (*N presa* +1)

CUADRO XV.4. *(Continuación)*

Capacidad de carga poblacional depredador (K_d)	100000 individuos
Crecimiento poblacional depredador	$\frac{N(K-N)}{K} \times e^{rt} - N$
Mortalidad depredador	0,1 x *N* depredador
Año	TIME
Capturas	0,001
Explotación	IF (Año>100) THEN (Capturas) ELSE(0)



BIBLIOGRAFÍA

Abraira V y Pérez de Vargas A (1996) *Métodos Multivariantes en estadística*. Editorial Centro de Estudios Ramón Areces, Madrid.

Aguilera P y Garrido A (2001) *Aplicaciones ambientales del análisis multivariable*. Universidad de Almería, Servicio de Publicaciones, Almería.

Akaike H (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723.

Azzimonti JC (2003) *Bioestadística aplicada a Bioquímica y Farmacia*. Universidad Nacional de Misiones. Editorial Universitaria, Argentina.

Box GEP y Jenkins JM (1976) *Time series analysis: Forecasting and control*. Holden-Day,San Francisco, CA.

Breslow NE y Day NE (1980) Statistical methods in cancer research. Volume 1. *The analysis or case-control studies*. Lyon: International Agency for Research on Cancer.

Callon M, Courtial JP y Penan H (1995) *Cienciometría: La medición de la actividad científica, de la bibliometría a la vigilancia tecnológic*a. Trea, D.L., Gijón.

Canavos GC (1993) *Probabilidad y Estadística - Aplicaciones y métodos*. McGraw-Hill, México DF.

Cochran WG (1954) Some methods for strengthening the common χ^2 tests. *Biometrics* 10: 417-451.

Cuadras CM (1991) *Métodos de Análisis Multivariante*. Promociones y Publicaciones Universitarias, Barcelona.

Davis JC (2002) *Statistic and data analysis in Geology*. John Wiley y Sons, New York.

Demènech I y Massons JM (1980) *Bioestadística. Métodos estadísticos para investigadores*. Ed. Herder. Barcelona.

Dubois DM (1973) An index of fluctuations, Do, connected with diversity and stability of ecosystems: applications in the Lotka–Volterra model and in an experimental distribution of species. *Rapport de sythèse III, Programme National sur l'environment Physique et Biologique, Project Mer*. Commision Interministérielle de la Politique Scientifique. Liège.

Erikson FH y Nosanchuk TA (1977) *Understandig data*. McGraw Hill. Toronto.

Fisher RA (1922) On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 85: 87-94.

Fisher LD y van Belle G (1993) *Biostatistics. A methodology for the health sciences.* John Wiley y Sons, New York.

Frangópulos M, Guisande C, deBlas E y Maneiro I (2004) Toxin production and competitive abilities under phosphorus limitation of *Alexandrium* species. *Harmful Algae* 3: 131-139.

Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32: 675–701.

Friis-Christensen E y Lassen K (1991) Length of the Solar Cycle: An Indicator of Solar Activity Closely Associated with Climate. *Science* 254: 698–700.

Gillman M y Hails R (1997) *An Introduction to Ecological Modelling*. Blackwell Science Ltd, Oxford.

Grimm V y Wissel C (1997) Babel, or the ecological stability discussions: an inventory and analysis of terminology and a guide for avoiding confusion. *Oecologia* 109: 323-334.

Hernández-Llamas A y Ratkowsky DA (2004) Growth of fishes, crustaceans and molluscs: estimation of the von Bertalanffy, Logistic, Gompertz and Richards curves and a new growth model. *Marine Ecology Progress Series* 282: 237-244.

Hollander M y Wolfe DA (1999) *Nonparametric Statistical Methods.* Ed. Wiley-Interscience, New York.

Horn HS (1966) Measured of "overlap" in comparative ecological studies. *American Naturalist* 100: 419-424.

Hyndman RJ y Khandakar Y (2008) Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software* 27: 1-22.

Isaaks EH y Srivastava RM (1989) *An Introduction to Applied Geostatistics*. Oxfrod University Press, Oxford.

Johnson RA y Wichern DW (1988) *Applied multivariate statistical analysis.* Prentice-Hall, Inc. Upper Saddle River, NJ, USA.

Kleinbaum DG, Dupper LL, Muller KE y Nizam A (1998) *Applied regression analysis and other multivariable methods*. Duxbury Press, California.

Krebs CHJ (1989) *Ecological methodology*. Harper Collins Publishers, New York.

Legendre P y Legendre L (1998) *Numerical Ecology*. *Developments in Environmental Modelling 20*. Elsevier, The Netherlands.

Lilliefors H (1967) On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* 62: 399–402.

Lotka AJ (1925) *Elements of Physical Biology.* William y Wilkins, Baltimore. Reimpreso en 1956 por Dover Publications, New York.

Lotka AJ (1927) Fluctuations in the abundance of species considered mathematically (with comment by V. Volterra). *Nature* 119**:** 12-13.

Mantel N y Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22: 719-748.

McNaughton SJ y Wolf LL (1984) *Ecología general*. Ediciones Omega, Barcelona.

McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12: 153–157.

Mardia KV, Kent JT y Bibby JM (1992). *Multivariate Analysis.* Academic Press. Londres

Moore DS (1998) *Estadística aplicada básica*. Antoni Bosch Editor.

Odum HT y Odum EC (2000) *Modelling for all Scales*. Academic Press, San Diego.

Peña D (1997) *Estadística. Modelos y métodos. 1. Fundamentos*. Alianza Universidad Textos, Madrid.

Peña D (2002) *Análisis de datos multivariantes*. McGraw-Hill /Interamericana de España. Madrid.

Peña D y Romo J (1999) *Introducción a la Estadística para las Ciencias Sociales*. McGraw-Hill, Madrid.

Pérez C (2004) *Técnica estadística con SPSS*. Universidad Complutense de Madrid. Prentice Hall, Madrid.

Pérez C (2004) *Técnicas de análisis multivariante de datos. Aplicaciones con SPSS*. Pearson Educación. Madrid.

Price RH y Boufard DL (1974) Behavioral appropriateness and situational constraint as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30: 579-586.

Quesada V, Isidoro A y López LA (1994) *Curso y ejercicios de Estadística. Aplicación a las Ciencias Biológicas, Médicas y Sociales*. Alhambra Longman, Madrid.

Reichel R, Thejll PA y Lassen K (2001) The cause-and-effect relationship of solar cycle length and the Northern hemisphere air surface temperature. *Journal of Geophysical Research - Space Physics* 106, No. A8: 15635-15641.

Salvarrey L (2000) *Curso de estadística básica*. Universidad de la República del Uruguay.

Sánchez JJ (1999) *Manual de análisis estadístico de los datos*. Alianza Editorial. Madrid.

Searle SR (1997) *Linear Models*. John Wiley y Sons (Wiley Classics Library). New York.

Siegel S y Castellan NJ Jr (1988) *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York.

Shapiro SS y Wilk MB (1965) An analysis of variance test for normality (complete samples). *Biometrika* 52: 591–611.

SigmaPlot (2002) Copyright 1986-2001 © SPSS Inc.

Smith RL y Smith TM (2000) *Ecología*. Pearson Educación, S.A., Madrid.

Sokal RR y Rohlf FJ (1979) *Biometría. Principios y métodos estadísticos en la investigación biológica. H. Blume Ediciones, Madrid.*

Sokal RR y Rohlf FJ (1980) *Introducción a la Bioestadística*. Editorial Reverté, Madrid.

Sokal RR y Rohlf FJ (1981) *Biometry*. WH Freeman and Company, New York.

Sprent P (1993) *Applied Nonparametric Statistical Methods*. Ed. Chapman y Hall, London.

SPSS (2003) Copyright 1989-2003 © SPSS Inc.

SPSS. Guía para el análisis de datos. http://www.spss.es.

STATISTICA 7 (2004) Copyright 1984-2006 © Statsoft Inc.

Steel RGD y Torrie JH (1985) *Bioestadística: Principios y procedimientos*. McGraw-Hill, Bogotá.

Stella (2001) Copyright 1985 – 2001 © High Performance Systems, Inc.

Tarone RE (1985) On heterogeneity tests based on efficient scores. *Biometrika* 72: 91-95.

Terradas J (2001) *Ecología de la vegetación*. Ediciones Omega, Barcelona.

Viedma JA (1989) *Bioestadística. Métodos en Medicina y Biología.* Librería Universitaria, Alicante.

Williams DA (1976) Improved llikelihood ratio test for complete contingency tables. *Biometrika* 63: 33-37.

Visauta B (2002) *Análisis estadístico con SPSS 11.0 para WINDOWS, Volumen I. Estadística básica.* McGraw Hill. Madrid.

Vizmanos JR y Asensio R (1976) *Bioestadística. Curso y ejercicios*. Centro de Promoción Reprográfica, Madrid.

Ruth M y Hannon BM (1997) *Modelling Dynamic Economic Systems*. Springer-Verlag, New York.

Volterra V (1926) Fluctuations in the abundance of species considered mathematically*. Nature* 118**:** 12-13.

Yates F (1934) Contingency table involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society* 1: 217-235.

Zar JH (1999) *Biostatistical Analysis*. Prentice Hall, New Jersey.

GUÍA RESUMEN

Esta guía tiene por objeto orientar al lector en la elección de la técnica aplicable. Se muestran algunos ejemplos de la operación estadística a realizar, el tipo de datos disponibles y el método o técnica más adecuado dentro de los que se describen en este libro. La lista no es en modo alguno exhaustiva; existen otras operaciones realizables con los métodos disponibles, y otras técnicas estadísticas aplicables para resolver los mismos problemas estadísticos, por lo que debe considerarse como una guía orientativa.

Tipo de medida u operación estadística	Tipo de datos	Método	Página
Representación de datos con más de dos variables		Coordenadas polares	1
Estandarización de datos			2
Medidas de posición no central		Percentiles, cuartiles y deciles	27
Medidas de posición central	Tasas de crecimiento	Media geométrica	22
	Velocidades con distancia constante, o precios con presupuesto constante	Media armónica	22
	Variables cualitativas en escala nominal	Moda	23
	Variables con distribución muy asimétrica	Mediana	23
	Variables con valores extremos	Media acotada	25
	Datos con diferente peso o importancia	Media ponderada	25
	Cualquier variable cuantitativa	Media aritmética (más frecuentemente usada)	21

Tipo de medida u operación estadística	Tipo de datos	Método	Página
Medidas de dispersión	Para todo tipo de datos	Amplitud	30
	Cuando se utiliza la media aritmética como medida de posición	Varianza o cuasivarianza	30
		Desviación típica o cuasidesviación típica (de uso más común porque se expresa en las mismas unidades que los datos)	30
	Para comparar la dispersión de la misma variable en poblaciones distintas	Coeficiente de variación	33
	Cuando se quiere medir la variabilidad de la media muestral (para estimar la media poblacional)	Error estándar de la media	33
	Se expresa la posición central por medio de la mediana	Recorrido intercuartílico	33
Calcular la probabilidad de que una variable tome valores en un determinado intervalo	Datos con distribución Normal	Estadístico *Z*	49
Calcular la probabilidad de uno o varios valores de una variable discreta	Los sucesos son independientes y con dos opciones complementarias y excluyentes (como cara y cruz)	Función de probabilidad binomial	62
	Idem con sucesos no independientes, ya que la muestra se extrae sin reposición de una población finita	Función de probabilidad hipergeométrica	64
	Sucesos independientes distribuidos al azar en el tiempo o en el espacio	Función de probabilidad de Poisson	66
Intervalo de confianza de la media de una población Normal	Desviación típica conocida	Estadístico *Z* con la desviación típica	71
	Desviación típica desconocida (se trabaja con una muestra del conjunto de toda la población) y $n \geq 30$	Estadístico *Z* con la cuasidesviación típica (aproximación de la *t*)	72
	Desviación típica desconocida (se trabaja con una muestra del conjunto de toda la población) y $n < 30$	Estadístico *t*	72

Tipo de medida u operación estadística	Tipo de datos	Método	Página
Intervalo de confianza de la diferencia de medias de dos poblaciones Normales	Varianzas conocidas	Estadístico *Z* con la desviación típica	83
	Varianza desconocida (una muestra del conjunto de toda la población) y n ≥ 30	Estadístico *Z* con la cuasidesviación típica (aproximación de la *t*)	83
	Varianza desconocida (se trabaja con una muestra del conjunto de toda la población) y n < 30	Estadístico *t*	83
Intervalo de confianza de la varianza y desviación típica de una población Normal	Variables Normales	Estadístico χ^2	91
Intervalo de confianza de la razón de varianzas de dos poblaciones Normales	Variables Normales	Estadístico *F*	94
Bondad de ajuste (determinación del tipo de distribución: Normal, Binomial, Poisson, etc.)	Variables categorizadas	Prueba χ^2	107
	Variables no categorizadas (variables cuantitativas, continuas o discretas, no agrupadas en intervalos o clases).	Test de Kolmogorov-Smirnov	108
	Contrastes específicos para la prueba de Normalidad	Test de Kolmogorov-Smirnov y Shapiro-Wilk	123
Pruebas de homogeneidad para variables cualitativas	Variables independientes y politómicas	Prueba χ^2	136
		Test de la G	136
	Variables independientes y dicotómicas	Prueba χ^2 con corrección de Yates	150
		Prueba de Fisher	150
	Variables ordinales relacionadas	ANOVA de Friedman	161
	Variables dependientes y dicotómicas	Prueba de McNemar	167
		Q de Cochran	179
Pruebas de homogeneidad para variables cuantitativas con distribución Normal y con homogeneidad de varianzas (métodos paramétricos)	Dos muestras independientes	*t*-test	218
	Dos muestras dependientes	*t*-test	240
	Dos o más muestras independientes	ANOVA	251

Tipo de medida u operación estadística	Tipo de datos	Método	Página
Continuación	Dos o más muestras independientes que están afectadas por otras variables	ANCOVA	289
	Dos o más muestras dependientes	ANOVA y ANCOVA con medidas repetidas	318
	Variables anidadas	ANOVA anidado	363
Pruebas de homogeneidad para variables cuantitativas sin distribución Normal o sin homogeneidad de varianzas (métodos no paramétricos)	Dos muestras independientes	Prueba de U de Mann-Whitney	386
		Test de rachas de Wald-Wolfowitz	387
		Prueba de Kolmogorov-Smirnov para dos muestras	388
	k muestras independientes	Contraste de la mediana	400
		ANOVA de Kruskal-Wallis	400
	Dos variables relacionadas o pareadas	Contraste de los signos	408
		Prueba de Wilcoxon para pares relacionados	408
	k variables relacionadas	ANOVA de Friedman	415
Pruebas de independencia en variables cualitativas	Variables politómicas	Chi-cuadrado de Pearson y test *G* de razón de verosimilitud	189
	Variables dicotómicas	Corrección de Yates y prueba de Fisher	202
		Pruebas de Cochran y Mantel-Haenszel	202
Pruebas de asociación en variables cualitativas	Variables politómicas	V de Cramer	189
		Coeficiente de contingencia	190
		Coeficiente de incertidumbre	190
	Variables dicotómicas	Phi	203
		Riesgo relativo, razón de ventajas (odds ratio), pruebas de Breslow-Day y de Tarone	203

Tipo de medida u operación estadística	Tipo de datos	Método	Página
Pruebas de independencia y asociación en variables cuantitativas	Dos variables supuestamente relacionadas	Coeficiente de correlación de Pearson	421
		Coeficiente de correlación de Spearman	422
		Coeficiente de Kendall	423
		Coeficiente Gamma	424
	Dos variables supuestamente relacionadas, controlando el efecto de otras variables	Coeficiente de correlación parcial de Pearson	439
		Coeficiente de correlación parcial de Kendall	440
Función que describe la relación entre variables cuantitativas	Una variable dependiente linealmente de otra	Regresión simple	455
	Una variable dependiente linealmente de otras	Regresión lineal múltiple	486
	Entre varias variables con relación no lineal	Regresión simple o múltiple no lineal	519
Función que describe la relación entre variables cualitativas	Variable dependiente dicotómica (con dos clases)	Regresión logística binomial	537
	Variable dependiente politómica (con más de dos clases)	Regresión logística multinomial	557
Series temporales	Construcción de modelos	Modelo ARIMA	585
	Ciclos	Análisis espectral	637
Métodos factoriales multivariantes. Reducir el número de variables para explicar los datos. Analizar las relaciones entre variables	Sin hipótesis previa. Método descriptivo. Variables cuantitativas correlacionadas.	Análisis de Componentes Principales	667
	Las variables dependen de factores comunes. Modelo explicativo. Variables cuantitativas correlacionadas	Análisis Factorial	697
	Variables cualitativas. Tablas grandes de frecuencias	Análisis de Correspondencias	719
	Encontrar la estructura subyacente en un conjunto de medidas de proximidad/distancia entre objetos. Variables cuantitativas y cualitativas	Escalamiento multidimensional	743

Tipo de medida u operación estadística	Tipo de datos	Método	Página
Clasificación de elementos o variables	Sin hipótesis a priori	Clasificación jerárquica	759
	Con idea previa del número de grupos	Clasificación de *k*-medias	773
Analizar si existen diferencias significativas entre grupos con respecto a un conjunto de variables. Asignar casos a grupos	Variables cuantitativas	Análisis Discriminante	787
Predecir el comportamiento de sistemas	Variables cuantitativas y cualitativas	Modelos de simulación	811

Apéndice I

TABLAS ESTADÍSTICAS

Tabla 1. Áreas bajo la curva Normal estándar a la izquierda de Z.

Primer decimal de Z	Segundo decimal de Z									
	0,00	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Tabla 2. Valores críticos de la distribución *t* de Student.

Grados de libertad (ν)	Nivel de significación unilateral (α)						
	0,25	**0,1**	**0,05**	**0,025**	**0,01**	**0,005**	**0,001**
1	1,000	3,078	6,314	12,706	31,821	63,657	318,289
2	0,816	1,886	2,920	4,303	6,965	9,925	22,328
3	0,765	1,638	2,353	3,182	4,541	5,841	10,214
4	0,741	1,533	2,132	2,776	3,747	4,604	7,173
5	0,727	1,476	2,015	2,571	3,365	4,032	5,894
6	0,718	1,440	1,943	2,447	3,143	3,707	5,208
7	0,711	1,415	1,895	2,365	2,998	3,499	4,785
8	0,706	1,397	1,860	2,306	2,896	3,355	4,501
9	0,703	1,383	1,833	2,262	2,821	3,250	4,297
10	0,700	1,372	1,812	2,228	2,764	3,169	4,144
11	0,697	1,363	1,796	2,201	2,718	3,106	4,025
12	0,695	1,356	1,782	2,179	2,681	3,055	3,930
13	0,694	1,35	1,771	2,160	2,650	3,012	3,852
14	0,692	1,345	1,761	2,145	2,624	2,977	3,787
15	0,691	1,341	1,753	2,131	2,602	2,947	3,733
16	0,690	1,337	1,746	2,120	2,583	2,921	3,686
17	0,698	1,333	1,740	2,110	2,567	2,898	3,646
18	0,688	1,330	1,734	2,101	2,552	2,878	3,610
19	0,688	1,328	1,729	2,093	2,539	2,861	3,579
20	0,687	1,325	1,725	2,086	2,528	2,845	3,552
21	0,686	1,323	1,721	2,080	2,518	2,831	3,527
22	0,686	1,321	1,717	2,074	2,508	2,819	3,505
23	0,685	1,319	1,714	2,069	2,500	2,807	3,485
24	0,685	1,318	1,711	2,064	2,492	2,797	3,467
25	0,684	1,316	1,708	2,060	2,485	2,787	3,450
26	0,684	1,315	1,706	2,056	2,479	2,779	3,435
27	0,684	1,314	1,703	2,052	2,473	2,771	3,421
28	0,683	1,313	1,701	2,048	2,467	2,763	3,408
29	0,683	1,311	1,699	2,045	2,462	2,756	3,396
30	0,683	1,310	1,697	2,042	2,457	2,750	3,385
40	0,681	1,303	1,684	2,021	2,423	2,704	3,307
60	0,679	1,296	1,671	2,000	2,390	2,660	3,232
120	0,677	1,289	1,658	1,980	2,358	2,617	3,160
∞	0,674	1,282	1,645	1,960	2,326	2,576	3,090

Tabla 3. Valores críticos de la distribución χ^2.

Grados de libertad (ν)	Nivel de significación (α)				
	0,995	**0,975**	**0,900**	**0,500**	**0,100**
1	0,000	0,000	0,016	0,455	2,706
2	0,010	0,051	0,211	1,386	4,605
3	0,072	0,216	0,584	2,366	6,251
4	0,207	0,484	1,064	3,357	7,779
5	0,412	0,831	1,610	4,351	9,236
6	0,676	1,237	2,204	5,348	10,645
7	0,989	1,690	2,833	6,346	12,017
8	1,344	2,180	3,490	7,344	13,362
9	1,735	2,700	4,168	8,343	14,684
10	2,156	3,247	4,865	9,342	15,987
11	2,603	3,816	5,578	10,341	17,275
12	3,074	4,404	6,304	11,340	18,549
13	3,565	5,009	7,042	12,340	19,812
14	4,075	5,629	7,790	13,339	21,064
15	4,601	6,262	8,547	14,339	22,307
16	5,142	6,908	9,312	15,338	23,542
17	5,697	7,564	10,085	16,338	24,769
18	6,265	8,231	10,865	17,338	25,989
19	6,844	8,907	11,651	18,338	27,204
20	7,434	9,591	12,443	19,337	28,412
21	8,034	10,283	13,240	20,337	29,615
22	8,643	10,982	14,042	21,337	30,813
23	9,260	11,688	14,848	22,337	32,007
24	9,886	12,401	15,659	23,337	33,196
25	10,520	13,120	16,473	24,337	34,382

Tabla 3. Valores críticos de la distribución χ^2 (*continuación*).

Grados de libertad (ν)	Nivel de significación (α)				
	0,050	**0,025**	**0,010**	**0,005**	**0,001**
1	3,841	5,024	6,635	7,879	10,828
2	5,991	7,378	9,210	10,597	13,816
3	7,815	9,348	11,345	12,838	16,266
4	9,488	11,143	13,277	14,860	18,467
5	11,070	12,832	15,086	16,750	20,515
6	12,592	14,449	16,812	18,548	22,458
7	14,067	16,013	18,475	20,278	24,322
8	15,507	17,535	20,090	21,955	26,124
9	16,919	19,023	21,666	23,589	27,877
10	18,307	20,483	23,209	25,188	29,588
11	19,675	21,920	24,725	26,757	31,264
12	21,026	23,337	26,217	28,300	32,910
13	22,362	24,736	27,688	29,819	34,528
14	23,685	26,119	29,141	31,319	36,123
15	24,996	27,488	30,578	32,801	37,697
16	26,296	28,845	32,000	34,267	39,252
17	27,587	30,191	33,409	35,718	40,790
18	28,869	31,526	34,805	37,156	42,312
19	30,144	32,852	36,191	38,582	43,820
20	31,410	34,170	37,566	39,997	45,315
21	32,670	35,479	38,932	41,401	46,797
22	33,924	36,781	40,289	42,796	48,268
23	35,172	38,076	41,638	44,181	49,728
24	36,415	39,364	42,980	45,558	51,179
25	37,652	40,646	44,314	46,928	52,620

Tabla 3. Valores críticos de la distribución χ^2 (*continuación*).

Grados de libertad (ν)	Nivel de significación (α)				
	0,995	**0,975**	**0,900**	**0,500**	**0,100**
26	11,160	13,844	17,292	25,336	35,563
27	11,808	14,573	18,114	26,336	36,741
28	12,461	15,308	18,939	27,336	37,916
29	13,121	16,047	19,768	28,336	39,088
30	13,787	16,791	20,599	29,336	40,256
31	14,458	17,539	21,434	30,336	41,422
32	15,134	18,291	22,271	31,336	42,585
33	15,815	19,047	23,110	32,336	43,745
34	16,501	19,806	23,952	33,336	44,903
35	17,192	20,569	24,797	34,336	46,059
36	17,887	21,336	25,643	35,336	47,212
37	18,586	22,106	26,492	36,335	48,363
38	19,289	22,878	27,343	37,335	49,513
39	19,996	23,654	28,196	38,335	50,660
40	20,707	24,433	29,051	39,335	51,805
41	21,421	25,215	29,907	40,335	52,949
42	22,138	25,999	30,765	41,335	54,090
43	22,859	26,785	31,625	42,335	55,230
44	23,584	27,575	32,487	43,335	56,369
45	24,311	28,366	33,350	44,335	57,505
46	25,042	29,160	34,215	45,335	58,641
47	25,775	29,956	35,081	46,335	59,774
48	26,511	30,755	35,949	47,335	60,907
49	27,249	31,555	36,818	48,335	62,038
50	27,991	32,357	37,689	49,335	63,167

Tabla 3. Valores críticos de la distribución χ^2 (*continuación*).

Grados de libertad (ν)	Nivel de significación (α)				
	0,050	**0,025**	**0,010**	**0,005**	**0,001**
26	38,885	41,923	45,642	48,290	54,052
27	40,113	43,194	46,963	49,645	55,476
28	41,337	44,461	48,278	50,993	56,892
29	42,557	45,722	49,588	52,336	58,301
30	43,773	46,979	50,892	53,672	59,703
31	44,985	48,232	52,191	55,003	61,098
32	46,194	49,480	53,486	56,329	62,487
33	47,400	50,725	54,776	57,649	63,870
34	48,602	51,966	56,061	58,964	65,247
35	49,802	53,203	57,342	60,275	66,619
36	50,998	54,437	58,619	61,582	67,985
37	52,192	55,668	59,892	62,884	69,346
38	53,384	56,896	61,162	64,182	70,703
39	54,572	58,120	62,428	65,476	72,055
40	55,758	59,342	63,691	66,766	73,402
41	56,942	60,561	64,950	68,053	74,745
42	58,124	61,777	66,206	69,336	76,084
43	59,304	62,990	67,459	70,616	77,419
44	60,481	64,202	68,710	71,893	78,750
45	61,656	65,410	69,957	73,166	80,077
46	62,830	66,617	71,201	74,437	81,400
47	64,001	67,821	72,443	75,704	82,720
48	65,171	69,023	73,683	76,969	84,037
49	66,339	70,222	74,919	78,231	85,351
50	67,505	71,420	76,154	79,490	86,661

Tabla 3. Valores críticos de la distribución χ^2 (*continuación*).

Grados de libertad (ν)	Nivel de significación (α)				
	0,995	**0,975**	**0,900**	**0,500**	**0,100**
51	28,735	33,162	38,560	50,335	64,295
52	29,481	33,968	39,433	51,335	65,422
53	30,230	34,776	40,308	52,335	66,548
54	30,981	35,586	41,183	53,335	67,673
55	31,735	36,398	42,060	54,335	68,796
56	32,490	37,212	42,937	55,335	69,918
57	33,248	38,027	43,816	56,335	71,040
58	34,008	38,844	44,696	57,335	72,160
59	34,770	39,662	45,577	58,335	73,279
60	35,534	40,482	46,459	59,335	74,397
61	36,300	41,303	47,342	60,335	75,514
62	37,068	42,126	48,226	61,335	76,630
63	37,838	42,950	49,111	62,335	77,745
64	38,610	43,776	49,996	63,335	78,860
65	39,383	44,603	50,883	64,335	79,973
66	40,158	45,431	51,770	65,335	81,085
67	40,935	46,261	52,659	66,335	82,197
68	41,713	47,092	53,548	67,334	83,308
69	42,494	47,924	54,438	68,334	84,418
70	43,275	48,758	55,329	69,334	85,527
71	44,058	49,592	56,221	70,334	86,635
72	44,843	50,428	57,113	71,334	87,743
73	45,629	51,265	58,006	72,334	88,850
74	46,417	52,103	58,900	73,334	89,956
75	47,206	52,942	59,795	74,334	91,061

Tabla 3. Valores críticos de la distribución χ^2 (*continuación*).

Grados de libertad (ν)	Nivel de significación (α)				
	0,050	**0,025**	**0,010**	**0,005**	**0,001**
51	68,669	72,616	77,386	80,747	87,968
52	69,832	73,810	78,616	82,001	89,272
53	70,993	75,002	79,843	83,253	90,573
54	72,153	76,192	81,069	84,502	91,872
55	73,311	77,380	82,292	85,749	93,168
56	74,468	78,567	83,513	86,994	94,460
57	75,624	79,752	84,733	88,237	95,751
58	76,778	80,936	85,950	89,477	97,039
59	77,931	82,117	87,166	90,715	98,324
60	79,082	83,298	88,379	91,952	99,607
61	80,232	84,476	89,591	93,186	100,888
62	81,381	85,654	90,802	94,419	102,166
63	82,529	86,830	92,010	95,649	103,442
64	83,675	88,004	93,217	96,878	104,716
65	84,821	89,177	94,422	98,105	105,988
66	85,965	90,349	95,626	99,331	107,258
67	87,108	91,519	96,828	100,550	108,526
68	88,250	92,689	98,028	101,780	109,791
69	89,391	93,856	99,228	103,000	111,055
70	90,531	95,023	100,430	104,210	112,317
71	91,670	96,189	101,620	105,430	113,577
72	92,808	97,353	102,820	106,650	114,835
73	93,945	98,516	104,010	107,860	116,092
74	95,081	99,678	105,200	109,070	117,346
75	96,217	100,840	106,390	110,290	118,599

Tabla 3. Valores críticos de la distribución χ^2 (*continuación*).

Grados de libertad (ν)	Nivel de significación (α)				
	0,995	**0,975**	**0,900**	**0,500**	**0,100**
76	47,997	53,782	60,690	75,334	92,166
77	48,788	54,623	61,586	76,334	93,270
78	49,582	55,466	62,483	77,334	94,373
79	50,376	56,309	63,380	78,334	95,476
80	51,172	57,153	64,278	79,334	96,578
81	51,969	57,998	65,176	80,334	97,680
82	52,767	58,845	66,076	81,334	98,780
83	53,567	59,692	66,976	82,334	99,880
84	54,368	60,540	67,876	83,334	100,980
85	55,170	61,389	68,777	84,334	102,080
86	55,973	62,239	69,679	85,334	103,180
87	56,777	63,089	70,581	86,334	104,280
88	57,582	63,941	71,484	87,334	105,370
89	58,389	64,793	72,387	88,334	106,470
90	59,196	65,647	73,291	89,334	107,560
91	60,005	66,501	74,196	90,334	108,660
92	60,815	67,356	75,101	91,334	109,760
93	61,625	68,211	76,006	92,334	110,850
94	62,437	69,068	76,912	93,334	111,940
95	63,250	69,925	77,818	94,334	113,040
96	64,063	70,783	78,725	95,334	114,130
97	64,878	71,642	79,633	96,334	115,220
98	65,694	72,501	80,541	97,334	116,320
99	66,510	73,361	81,449	98,334	117,410
100	67,328	74,222	82,358	99,334	118,500

Tabla 3. Valores críticos de la distribución χ^2 (*continuación*).

Grados de libertad (ν)	Nivel de significación (α)				
	0,050	**0,025**	**0,010**	**0,005**	**0,001**
76	97,351	102,000	107,580	111,500	119,850
77	98,484	103,160	108,770	112,700	121,100
78	99,617	104,320	109,960	113,910	122,348
79	100,750	105,470	111,140	115,120	123,594
80	101,880	106,630	112,330	116,320	124,839
81	103,010	107,780	113,510	117,520	126,082
82	104,140	108,940	114,690	118,730	127,324
83	105,270	110,090	115,880	119,930	128,565
84	106,390	111,240	117,060	121,130	129,804
85	107,520	112,390	118,240	122,320	131,041
86	108,650	113,540	119,410	123,520	132,277
87	109,770	114,690	120,590	124,720	133,512
88	110,900	115,840	121,770	125,910	134,745
89	112,020	116,990	122,940	127,110	135,978
90	113,150	118,140	124,120	128,300	137,208
91	114,270	119,280	125,290	129,490	138,438
92	115,390	120,430	126,460	130,680	139,666
93	116,510	121,570	127,630	131,870	140,893
94	117,630	122,720	128,800	133,060	142,119
95	118,750	123,860	129,970	134,250	143,344
96	119,870	125,000	131,140	135,430	144,567
97	120,990	126,140	132,310	136,620	145,789
98	122,110	127,280	133,480	137,800	147,010
99	123,230	128,420	134,640	138,990	148,230
100	124,340	129,560	135,810	140,170	149,449

Tabla 4. Valores críticos de la distribución *F* de Fisher-Snedecor para un α de 0,01. Grados de libertad del numerador en columnas y del denominador en filas.

υ/ω	1	2	3	4	5	6	7	8	9
1	4052,19	4999,52	5403,34	5624,62	5763,65	5858,97	5928,33	5981,1	6022,5
2	98,502	99,000	99,166	99,249	99,300	99,333	99,356	99,374	99,388
3	34,116	30,816	29,457	28,710	28,237	27,911	27,672	27,489	27,345
4	21,198	18,000	16,694	15,977	15,522	15,207	14,976	14,799	14,659
5	16,258	13,274	12,060	11,392	10,967	10,672	10,456	10,289	10,158
6	13,745	10,925	9,780	9,148	8,746	8,466	8,260	8,102	7,976
7	12,246	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719
8	11,259	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911
9	10,561	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,351
10	10,044	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744	4,632
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,388
13	9,074	6,701	5,739	5,205	4,862	4,620	4,441	4,302	4,191
14	8,862	6,515	5,564	5,035	4,695	4,456	4,278	4,140	4,030
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,895
16	8,531	6,226	5,292	4,773	4,437	4,202	4,026	3,890	3,780
17	8,400	6,112	5,185	4,669	4,336	4,102	3,927	3,791	3,682

Tabla 4. Valores críticos de la distribución *F* de Fisher-Snedecor para un α de 0,01 (*continuación*).

υ/ω	10	12	15	20	24	30	40	60	120	∞
1	6055,85	6106,35	6157,28	6208,74	6234,6	6260,6	6286,8	6313	6339,4	6365,8
2	99,399	99,416	99,432	99,449	99,458	99,466	99,474	99,482	99,491	99,499
3	27,229	27,052	26,872	26,690	26,598	26,505	26,411	26,316	26,221	26,125
4	14,546	14,374	14,198	14,020	13,929	13,838	13,745	13,652	13,558	13,463
5	10,051	9,888	9,722	9,553	9,466	9,379	9,291	9,202	9,112	9,020
6	7,874	7,718	7,559	7,396	7,313	7,229	7,143	7,057	6,969	6,880
7	6,620	6,469	6,314	6,155	6,074	5,992	5,908	5,824	5,737	5,650
8	5,814	5,667	5,515	5,359	5,279	5,198	5,116	5,032	4,946	4,859
9	5,257	5,111	4,962	4,808	4,729	4,649	4,567	4,483	4,398	4,311
10	4,849	4,706	4,558	4,405	4,327	4,247	4,165	4,082	3,996	3,909
11	4,539	4,397	4,251	4,099	4,021	3,941	3,860	3,776	3,690	3,602
12	4,296	4,155	4,010	3,858	3,780	3,701	3,619	3,535	3,449	3,361
13	4,100	3,960	3,815	3,665	3,587	3,507	3,425	3,341	3,255	3,165
14	3,939	3,800	3,656	3,505	3,427	3,348	3,266	3,181	3,094	3,004
15	3,805	3,666	3,522	3,372	3,294	3,214	3,132	3,047	2,959	2,868
16	3,691	3,553	3,409	3,259	3,181	3,101	3,018	2,933	2,845	2,753
17	3,593	3,455	3,312	3,162	3,084	3,003	2,920	2,835	2,746	2,653

Tabla 4. Valores críticos de la distribución *F* de Fisher-Snedecor para un α de 0,01 (*continuación*).

υ/ω	1	2	3	4	5	6	7	8	9
18	8,285	6,013	5,092	4,579	4,248	4,015	3,841	3,705	3,597
19	8,185	5,926	5,010	4,500	4,171	3,939	3,765	3,631	3,523
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,457
21	8,017	5,780	4,874	4,369	4,042	3,812	3,640	3,506	3,398
22	7,945	5,719	4,817	4,313	3,988	3,758	3,587	3,453	3,346
23	7,881	5,664	4,765	4,264	3,939	3,710	3,539	3,406	3,299
24	7,823	5,614	4,718	4,218	3,895	3,667	3,496	3,363	3,256
25	7,770	5,568	4,675	4,177	3,855	3,627	3,457	3,324	3,217
26	7,721	5,526	4,637	4,140	3,818	3,591	3,421	3,288	3,182
27	7,677	5,488	4,601	4,106	3,785	3,558	3,388	3,256	3,149
28	7,636	5,453	4,568	4,074	3,754	3,528	3,358	3,226	3,120
29	7,598	5,420	4,538	4,045	3,725	3,499	3,330	3,198	3,092
30	7,562	5,390	4,510	4,018	3,699	3,473	3,305	3,173	3,067
40	7,314	5,179	4,313	3,828	3,514	3,291	3,124	2,993	2,888
60	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823	2,718
120	6,851	4,787	3,949	3,480	3,174	2,956	2,792	2,663	2,559
∞	6,635	4,605	3,782	3,319	3,017	2,802	2,639	2,511	2,407

Tabla 4. Valores críticos de la distribución *F* de Fisher-Snedecor para un α de 0,01 (*continuación*).

υ/ω	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
18	3,508	3,371	3,227	3,077	2,999	2,919	2,835	2,749	2,660	2,566
19	3,434	3,297	3,153	3,003	2,925	2,844	2,761	2,674	2,584	2,489
20	3,368	3,231	3,088	2,938	2,859	2,778	2,695	2,608	2,517	2,421
21	3,310	3,173	3,030	2,880	2,801	2,720	2,635	2,548	2,457	2,360
22	3,258	3,121	2,978	2,827	2,749	2,667	2,583	2,495	2,403	2,305
23	3,211	3,074	2,931	2,781	2,702	2,620	2,535	2,447	2,354	2,256
24	3,168	3,032	2,889	2,738	2,659	2,577	2,492	2,403	2,310	2,211
25	3,129	2,993	2,850	2,699	2,620	2,538	2,453	2,364	2,270	2,169
26	3,094	2,958	2,815	2,664	2,585	2,503	2,417	2,327	2,233	2,131
27	3,062	2,926	2,783	2,632	2,552	2,470	2,384	2,294	2,198	2,097
28	3,032	2,896	2,753	2,602	2,522	2,440	2,354	2,263	2,167	2,064
29	3,005	2,868	2,726	2,574	2,495	2,412	2,325	2,234	2,138	2,034
30	2,979	2,843	2,700	2,549	2,469	2,386	2,299	2,208	2,111	2,006
40	2,801	2,665	2,522	2,369	2,288	2,203	2,114	2,019	1,917	1,805
60	2,632	2,496	2,352	2,198	2,115	2,028	1,936	1,836	1,726	1,601
120	2,474	2,336	2,192	2,035	1,950	1,860	1,763	1,656	1,533	1,381
∞	2,321	2,185	2,039	1,878	1,791	1,696	1,592	1,473	1,325	1,000

Tabla 4. Valores críticos de la distribución *F* de Fisher-Snedecor para un α de 0,025.

υ/ω	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	648	799	864	900	922	937	948	957	963
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98

Tabla 4. Valores críticos de la distribución *F* de Fisher-Snedecor para un α de 0,025 (*continuación*).

υ/ω	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	969	977	985	993	997	1001	1006	1010	1014	1018
2	39,40	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50
3	14,42	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95	13,90
4	8,84	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
5	6,62	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
6	5,46	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
7	4,76	4,67	4,57	4,47	4,41	4,36	4,31	4,25	4,20	4,14
8	4,30	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
9	3,96	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
10	3,72	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
11	3,53	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
12	3,37	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,73
13	3,25	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
14	3,15	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
15	3,06	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
16	2,99	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
17	2,92	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25

Tabla 4. Valores críticos de la distribución *F* de Fisher-Snedecor para un α de 0,025 (*continuación*).

υ/ω	1	2	3	4	5	6	7	8	9
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22
∞	3,84	3,69	3,12	2,8	2,57	2,41	2,29	2,19	2,11

Tabla 4. Valores críticos de la distribución *F* de Fisher-Snedecor para un α de 0,025 (*continuación*).

υ/ω	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
18	2,87	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
19	2,82	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
20	2,77	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
21	2,73	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
22	2,70	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
23	2,67	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
24	2,64	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
25	2,61	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
26	2,59	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
27	2,57	2,47	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
28	2,55	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
29	2,53	2,43	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
30	2,51	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
40	2,39	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
60	2,27	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
120	2,16	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
∞	2,05	1,95	1,83	1,71	1,64	1,57	1,49	1,39	1,27	1,00

Tabla 4. Valores críticos de la distribución *F* de Fisher-Snedecor para un α de 0,05.

υ/ω	1	2	3	4	5	6	7	8	9
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49

Tabla 4. Valores críticos de la distribución *F* de Fisher-Snedecor para un α de 0,05 (*continuación*).

υ/ω	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
1	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
6	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,41
12	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96

Tabla 4. Valores críticos de la distribución *F* de Fisher-Snedecor para un α de 0,05 (*continuación*).

υ/ω	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96
∞	3,84	3,00	2,61	2,4	2,21	2,10	2,01	1,94	1,88

Tabla 4. Valores críticos de la distribución *F* de Fisher-Snedecor para un α de 0,05 (*continuación*).

υ/ω	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
18	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,26
∞	1,83	1,75	1,67	1,57	1,52	1,46	1,40	1,32	1,22	1,00

Tabla 5. Valores críticos de la distribución del estadístico de Kolmogorov-Smirnov.

Grados de libertad (ν)	Nivel de significación (α)				
	0,2	**0,1**	**0,05**	**0,02**	**0,01**
1	0,900	0,950	0,975	0,990	0,995
2	0,684	0,776	0,842	0,900	0,929
3	0,565	0,636	0,708	0,785	0,829
4	0,493	0,565	0,624	0,689	0,734
5	0,447	0,509	0,563	0,627	0,669
6	0,410	0,468	0,519	0,577	0,617
7	0,381	0,436	0,483	0,538	0,576
8	0,358	0,410	0,454	0,507	0,542
9	0,339	0,387	0,430	0,480	0,513
10	0,323	0,369	0,409	0,457	0,489
11	0,308	0,352	0,391	0,437	0,468
12	0,296	0,338	0,375	0,419	0,449
13	0,285	0,325	0,361	0,404	0,432
14	0,275	0,314	0,349	0,390	0,418
15	0,266	0,304	0,338	0,377	0,404
16	0,258	0,295	0,327	0,366	0,392
17	0,250	0,286	0,318	0,355	0,381
18	0,244	0,279	0,309	0,346	0,371
19	0,237	0,271	0,301	0,337	0,361
20	0,232	0,265	0,294	0,329	0,352
21	0,226	0,259	0,287	0,321	0,344
22	0,221	0,253	0,281	0,314	0,337
23	0,216	0,247	0,275	0,307	0,330
24	0,212	0,242	0,269	0,301	0,323
25	0,208	0,238	0,264	0,295	0,317
26	0,204	0,233	0,259	0,290	0,311
27	0,200	0,229	0,254	0,284	0,305
28	0,197	0,225	0,250	0,279	0,300
29	0,193	0,221	0,246	0,275	0,295
30	0,190	0,218	0,242	0,270	0,290
31	0,187	0,214	0,238	0,266	0,285
32	0,184	0,211	0,234	0,262	0,281
33	0,182	0,208	0,231	0,258	0,277
34	0,179	0,205	0,227	0,254	0,273
35	0,177	0,202	0,224	0,251	0,269
36	0,174	0,199	0,221	0,247	0,265
37	0,172	0,196	0,218	0,244	0,262
38	0,170	0,194	0,215	0,241	0,258
39	0,168	0,191	0,213	0,238	0,255
40	0,165	0,189	0,210	0,235	0,252
Más de 40	$\frac{1,07}{\sqrt{n}}$	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,52}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

Tabla 6. Valores críticos de la distribución del estadístico de Lilliefors.

Grados de libertad (ν)	Nivel de significación (α)				
	0,2	**0,15**	**0,10**	**0,05**	**0,01**
4	0,300	0,319	0,352	0,381	0,417
5	0,285	0,299	0,315	0,337	0,405
6	0,265	0,277	0,294	0,319	0,364
7	0,247	0,258	0,276	0,300	0,348
8	0,233	0,244	0,261	0,285	0,331
9	0,223	0,233	0,249	0,271	0,311
10	0,215	0,224	0,239	0,258	0,294
11	0,206	0,217	0,230	0,249	0,284
12	0,199	0,212	0,223	0,242	0,275
13	0,190	0,202	0,214	0,234	0,268
14	0,183	0,194	0,207	0,227	0,261
15	0,177	0,187	0,201	0,220	0,257
16	0,173	0,182	0,195	0,213	0,250
17	0,169	0,177	0,189	0,206	0,245
18	0,166	0,173	0,184	0,200	0,239
19	0,163	0,169	0,179	0,195	0,235
20	0,160	0,166	0,174	0,190	0,231
25	0,149	0,153	0,165	0,180	0,203
30	0,131	0,136	0,144	0,161	0,187
Más de 30	$\frac{0,736}{\sqrt{n}}$	$\frac{0,768}{\sqrt{n}}$	$\frac{0,805}{\sqrt{n}}$	$\frac{0,886}{\sqrt{n}}$	$\frac{1,031}{\sqrt{n}}$

Tabla 7, Valores críticos de la distribución del estadístico del contraste Durbin-Watson. Se tabulan los valores de d_L y d_U para un nivel de significación de α = 0,05. El valor de *k* indica el número de variables independientes de la regresión.

	***k* = 1**		***k* = 2**		***k* = 3**		***k* = 4**		***k* = 5**	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
10	0,879	1,319	0,697	1,641	0,525	2,016	0,376	2,414	0,243	2,822
11	0,927	1,324	0,758	1,604	0,595	1,928	0,444	2,283	0,315	2,644
12	0,971	1,331	0,812	1,579	0,658	1,864	0,512	2,177	0,379	2,506
13	1,010	1,340	0,861	1,562	0,714	1,816	0,574	2,094	0,444	2,390
14	1,045	1,350	0,905	1,551	0,767	1,779	0,632	2,029	0,505	2,296
15	1,077	1,361	0,945	1,543	0,814	1,750	0,685	1,977	0,562	2,220
16	1,106	1,371	0,982	1,539	0,857	1,728	0,734	1,935	0,615	2,157
17	1,133	1,381	1,015	1,536	0,897	1,710	0,779	1,900	0,664	2,104
18	1,158	1,392	1,046	1,535	0,933	1,696	0,820	1,872	0,710	2,060
19	1,180	1,401	1,075	1,535	0,967	1,685	0,859	1,848	0,752	2,022
20	1,201	1,411	1,100	1,537	0,998	1,676	0,894	1,828	0,792	1,991
21	1,221	1,420	1,125	1,538	1,026	1,669	0,927	1,812	0,828	1,964
22	1,240	1,429	1,147	1,541	1,053	1,664	0,958	1,797	0,863	1,940
23	1,257	1,437	1,168	1,543	1,078	1,660	0,986	1,786	0,895	1,919
24	1,273	1,446	1,188	1,546	1,101	1,657	1,013	1,775	0,925	1,902
25	1,288	1,454	1,206	1,550	1,123	1,654	1,038	1,767	0,953	1,886
26	1,302	1,461	1,224	1,553	1,143	1,652	1,062	1,759	0,979	1,873
27	1,316	1,468	1,240	1,556	1,162	1,651	1,083	1,753	1,004	1,861
28	1,328	1,476	1,255	1,560	1,181	1,650	1,104	1,747	1,028	1,850
29	1,341	1,483	1,270	1,563	1,198	1,650	1,124	1,743	1,050	1,841
30	1,352	1,489	1,284	1,567	1,214	1,650	1,143	1,739	1,070	1,833
31	1,363	1,496	1,297	1,570	1,229	1,650	1,160	1,735	1,090	1,825
32	1,373	1,502	1,309	1,573	1,244	1,650	1,177	1,732	1,109	1,819
33	1,383	1,508	1,321	1,577	1,258	1,651	1,193	1,730	1,127	1,813
34	1,393	1,514	1,332	1,580	1,271	1,652	1,208	1,728	1,144	1,807
35	1,402	1,519	1,343	1,584	1,283	1,653	1,222	1,726	1,160	1,803
36	1,411	1,524	1,354	1,587	1,295	1,654	1,236	1,725	1,175	1,799
37	1,419	1,530	1,364	1,590	1,307	1,655	1,249	1,723	1,190	1,795
38	1,427	1,535	1,373	1,594	1,317	1,656	1,261	1,723	1,204	1,792
39	1,435	1,540	1,382	1,597	1,328	1,658	1,273	1,722	1,218	1,789
40	1,442	1,544	1,391	1,600	1,338	1,659	1,285	1,721	1,231	1,786
45	1,475	1,566	1,430	1,615	1,383	1,666	1,336	1,720	1,287	1,776
50	1,503	1,585	1,462	1,628	1,421	1,674	1,378	1,721	1,334	1,771
55	1,527	1,601	1,490	1,640	1,452	1,681	1,414	1,724	1,374	1,768
60	1,549	1,616	1,514	1,652	1,480	1,689	1,444	1,727	1,408	1,767
65	1,567	1,629	1,536	1,662	1,503	1,696	1,471	1,731	1,438	1,767
70	1,583	1,641	1,554	1,671	1,524	1,703	1,494	1,735	1,464	1,768
75	1,598	1,652	1,571	1,680	1,543	1,709	1,515	1,739	1,486	1,770
80	1,611	1,662	1,586	1,688	1,560	1,715	1,534	1,743	1,507	1,772
85	1,624	1,671	1,600	1,696	1,575	1,721	1,551	1,747	1,525	1,774
90	1,635	1,679	1,612	1,703	1,589	1,726	1,566	1,751	1,542	1,776
95	1,645	1,687	1,623	1,709	1,602	1,732	1,579	1,755	1,557	1,778
100	1,654	1,694	1,634	1,715	1,613	1,736	1,592	1,758	1,571	1,780
150	1,720	1,747	1,706	1,760	1,693	1,774	1,679	1,788	1,665	1,802
200	1,758	1,779	1,748	1,789	1,738	1,799	1,728	1,809	1,718	1,820
300	1,804	1,817	1,797	1,824	1,790	1,831	1,784	1,838	1,777	1,845

APÉNDICE II: INSTALACIÓN Y CONCEPTOS BÁSICOS SOBRE R

El objetivo de este apéndice es ayudar al lector a familiarizarse con aspectos básicos de R. De esta manera se complementa el contenido de los capítulos en los cuales se suministra una serie de scripts programados para la realización de cada una de las técnicas estadísticas descritas.

Estos scripts están desarrollados para que cualquier usuario sin ningún conocimiento de R pueda utilizarlos. No obstante, resulta muy recomendable, para cualquier persona interesada por el contenido de este libro, iniciarse en el manejo de este lenguaje de programación.

AII.1. Sobre R

AII.1.1. ¿QUÉ ES R?

R es un lenguaje de programación para estadística y matemáticas desarrollado inicialmente por Ihaka y Gentlemen en 1997. Toma como base las características del software comercial S-PLUS. R es de código abierto y distribución gratuita. Dispone de una serie de facilidades para manejo de datos, cálculos, almacenamiento, despliegue gráfico, etc.

Su popularidad es cada vez mayor entre la comunidad científica. Su éxito se debe principalmente a las siguientes razones: estabilidad, bajo consumo de recursos, dinamismo y gratuidad.

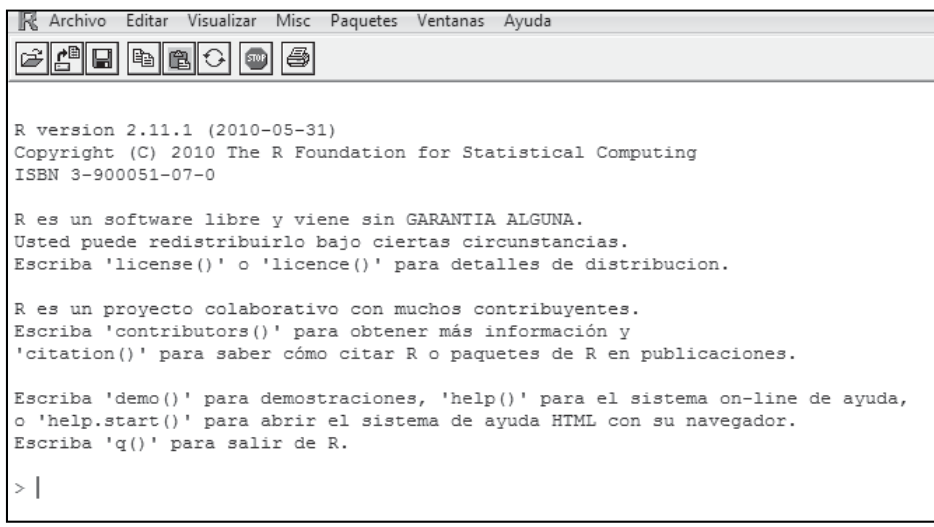
Decimos que R es dinámico porque, en primer lugar, permite a cualquier usuario programar nuevas aplicaciones (paquetes) y ponerlas a disposición de la comunidad de usuarios. Por lo tanto, R evoluciona y se desarrolla de forma continua apoyado en una gran comunidad de usuarios que programan y estudian el funcionamiento de sus aplicaciones. En segundo lugar, la interacción usuario/software es intuitiva y directa, lo que permite mucho control sobre cada paso. Actualmente, su desarrollo corre a cargo del *R Development Core Team*.

AII.1.2. INSTALACIÓN

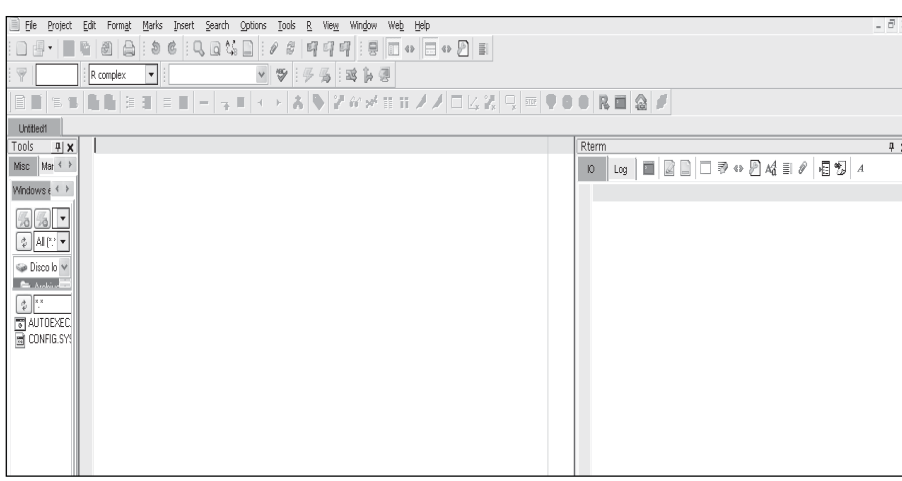
Podemos encontrar la página web de R en la dirección http://cran.r-project.org/. Allí veremos, en la parte superior, un enlace para la instalación de R en Linux, MacOS y Windows. Para Windows debemos seguir el enlace al subdirectorio *base*. En el caso de MacOS, el enlace nos lleva a una página donde podemos encontrar el archivo de instalación con la última versión. Hay varias formas de instalar R en distribuciones de Linux. De forma general, primero es necesario añadir «CRAN» a la lista de repositorios, configurar el nuevo repositorio adecuadamente e instalar el r-base.

AII.1.3. INTERFAZ

Una vez instalado, R se puede utilizar desde una consola, con una interfaz de línea de instrucciones. Es el método de mayor fiabilidad y estabilidad, pero poco atractivo para muchos usuarios. Como alternativa existen diferentes entornos gráficos que facilitan la interacción con el usuario (llamados GUI, del inglés *Graphical Users Interface*). Entre los más populares se encuentran *RGui* (disponible para Windows), *Tinn-R* (Windows, MacOS), *R Commander* (Windows, MacOS, Linux), *Emacs* (Windows, MacOS, Linux), *RKward* (Linux), pero existen muchos otros. En este libro hemos elaborado la mayoría de los ejemplos con *RGui* en Windows, y algunos con *R Commander*. La siguiente pantalla muestra el aspecto de la interfaz *RGui*, disponible para Windows:


Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
R version 2.11.1 (2010-05-31)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
>

Aspecto de la interfaz de *Tinn-R*, disponible para Windows y MacOs:



RGui se instala de forma automática con el paquete *base* de Windows. El resto de GUIs es necesario instalarlas específicamente. Se encuentran gratuitamente en internet.

Estas GUIs, para facilitar la interacción con el usuario, disponen de una serie de menús desplegables, cuya complejidad puede ir desde solamente algunas sencillas posibilidades de edición, apertura de ficheros, guardar datos, cargar paquetes, actualizar, etc. como podría ser el caso de *RGui*, hasta corrección de sintaxis de instrucciones, menús desplegables para importar/exportar datos, gráficos y múltiples técnicas estadísticas, como es el caso de *R Commander* y *RKward*.

AII.2. Primeros pasos

AII.2.1. ENTRAR/SALIR DE R Y SELECCIONAR DIRECTORIO DE TRABAJO

Si queremos entrar en R desde una consola, para trabajar con una interfaz de línea de instrucciones, simplemente deberemos teclear R en una línea. Si disponemos de alguno de los múltiples tipos de interfaz gráfica, además de a través de la consola, podemos usar el correspondiente icono de acceso directo.

Una vez dentro del programa, es muy importante seleccionar el directorio de trabajo. Este directorio es el que estará disponible de forma automática para cargar archivos de datos, abrir scripts, guardar áreas de trabajo y archivos de salida, etc. Si no seleccionamos un directorio de trabajo específico, nos dejará el que tiene por defecto, que depende de la instalación que se haya hecho. El directorio de trabajo se selecciona con la siguiente instrucción:

```
> setwd("C:/Documents and Settings/All Users/ArchivosR")
```

La función «setwd()» permite seleccionar el directorio deseado escribiendo como argumento toda la ruta de directorios entrecomillada, con las barras separadoras orientadas hacia la derecha. Si queremos saber cuál es el directorio de trabajo actual, debemos escribir la siguiente instrucción:

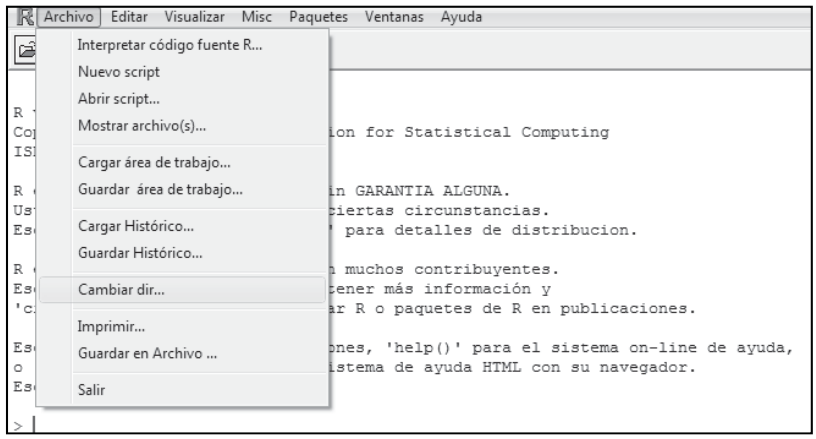
```
> getwd()
```

Aquí tenemos un ejemplo de cómo establecer un directorio de trabajo por instrucciones y después preguntar cuál es el directorio de trabajo actual:

```
> setwd("C:/Archivos de programa")
> getwd()
```

```
> setwd("C:/Archivos de programa")
> getwd()
[1] "C:/Archivos de programa"
```

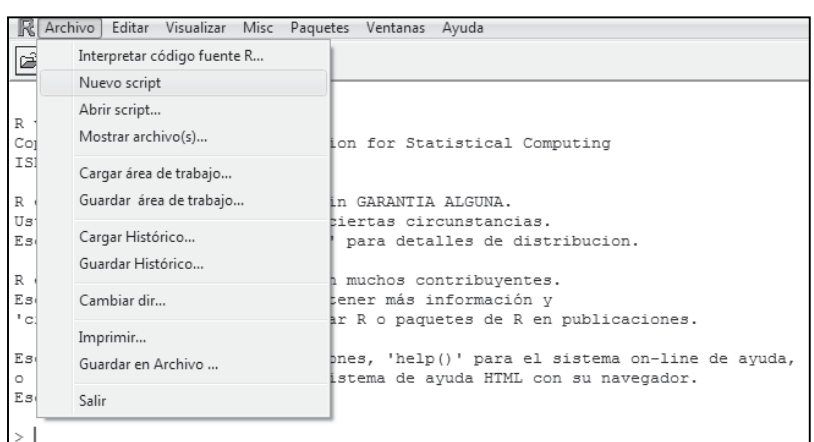
La mayoría de las GUIs disponen de una opción a través de menú para establecer el directorio de trabajo. En *RGui*, por ejemplo, se encuentra en el menú «Archivo» ⇨ «Cambiar dir…». Tras pulsar en esta opción surge un cuadro de diálogo en el que seleccionaremos nuestro directorio.



Para salir de R es suficiente con teclear «q()» en la línea de instrucciones. Si trabajamos con cualquiera de los GUIs, suele haber además un menú con la opción «Salir» (como se observa en la anterior ventana). Cualquiera que sea la forma que escojamos para salir, se nos preguntará si queremos guardar o no el área de trabajo. En el área de trabajo se guardan objetos que hayamos creado y que quedan disponibles hasta la próxima vez que deseemos utilizarlos abriendo el archivo. Se guarda como un archivo con extensión **Rdata**. Es posible tener abiertas simultáneamente varias áreas de trabajo. También es posible guardar la serie de instrucciones que hayamos escrito, como archivo **Rhistory**. No obstante, para esto es más interesante la opción de editarlos creando un script, como veremos a continuación.

AII.2.2. CREAR UN ARCHIVO SCRIPT

En las GUIs suele haber algún editor de texto en el que se puede crear lo que se llama habitualmente script (que en inglés puede significar *guión*). Mostraremos cómo se hace utilizando el RGui de Windows. Para crear un archivo donde se pueden insertar instrucciones para realizar determinadas funciones con R, simplemente hay que ir a «Archivo» ⇨ «Nuevo Script», como se muestra en la siguiente ventana.

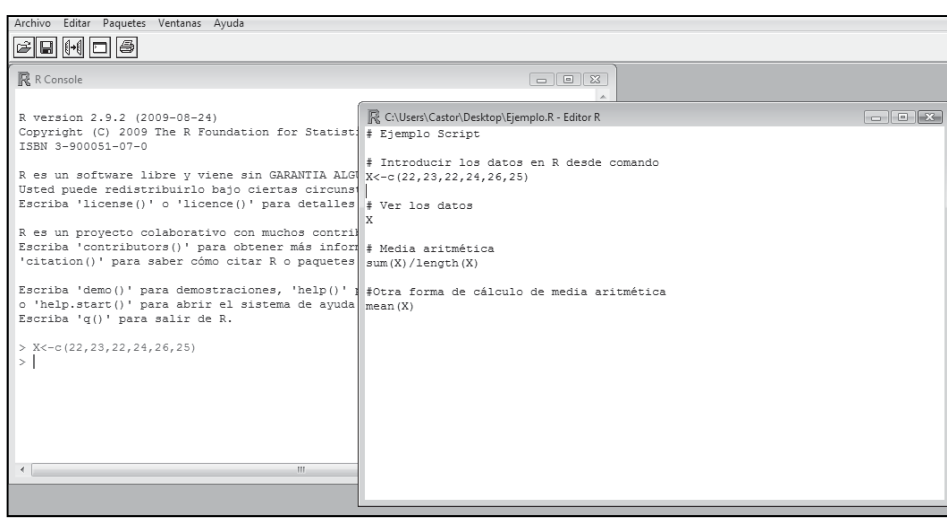


Una vez que se abre el archivo se puede comenzar a escribir las instrucciones. Las líneas que comienzan con el símbolo «#» son las correspondientes a comentario y no son ejecutadas como instrucciones de R.

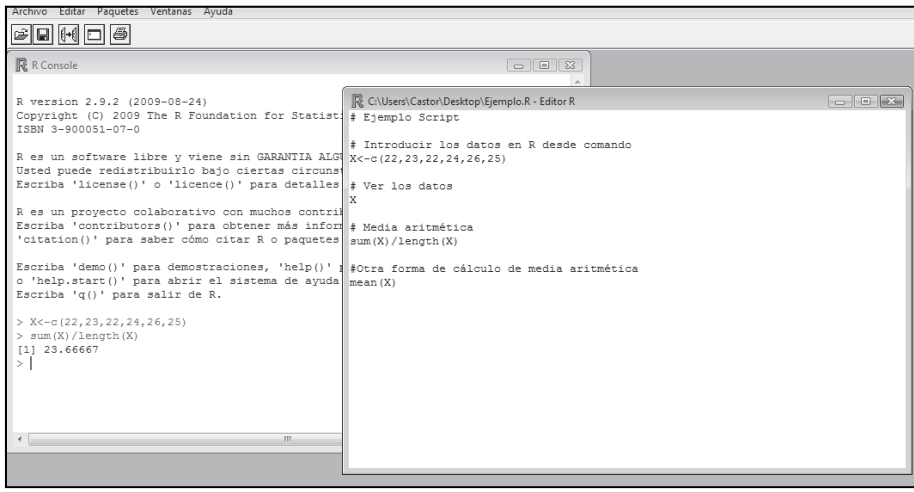
```
# Ejemplo Script

# Introducir los datos en R desde una instrucción
X<-c(22,23,22,24,26,25)

# Ver los datos
X

# Media aritmética
sum(X)/length(X)

# Otra forma
mean(X)
```

Se puede ir ejecutando línea por línea al colocar el cursor sobre la línea que se quiere ejecutar y se presiona la combinación *ctrl-R*. El comando pasa a la consola de R y se ejecuta.

En primer lugar hay que cargar los datos posicionándonos sobre la línea de introducción de datos y pulsando *ctrl-R*. En la medida que se envía cada instrucción se puede ir observando el resultado en la consola de R.


Archivo Editar Paquetes Ventanas Ayuda
R Console
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statist
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALG
Usted puede redistribuirlo bajo ciertas circuns
Escriba 'license()' o 'licence()' para detalles
R es un proyecto colaborativo con muchos contri
Escriba 'contributors()' para obtener más infor
'citation()' para saber cómo citar R o paquetes
Escriba 'demo()' para demostraciones, 'help()'
o 'help.start()' para abrir el sistema de ayuda
Escriba 'q()' para salir de R.
> X<-c(22,23,22,24,26,25)
>
C:\Users\Castor\Desktop\Ejemplo.R - Editor R
# Ejemplo Script
# Introducir los datos en R desde comando
X<-c(22,23,22,24,26,25)
# Ver los datos
X
# Media aritmética
sum(X)/length(X)
#Otra forma de cálculo de media aritmética
mean(X)

Si repetimos la operación pero en este caso posicionamos el cursor en media aritmética, obtenemos el resultado de la media en la consola de R como muestra la siguiente ventana. De la misma forma se puede posicionar en cualquier comando y ejecutarlo con *ctrl-R*.


Archivo Editar Paquetes Ventanas Ayuda
R Console
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statisti
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA ALG
Usted puede redistribuirlo bajo ciertas circuns
Escriba 'license()' o 'licence()' para detalles
R es un proyecto colaborativo con muchos contri
Escriba 'contributors()' para obtener más infor
'citation()' para saber cómo citar R o paquetes
Escriba 'demo()' para demostraciones, 'help()'
o 'help.start()' para abrir el sistema de ayuda
Escriba 'q()' para salir de R.
> X<-c(22,23,22,24,26,25)
> sum(X)/length(X)
[1] 23.66667
>
C:\Users\Castor\Desktop\Ejemplo.R - Editor R
# Ejemplo Script
# Introducir los datos en R desde comando
X<-c(22,23,22,24,26,25)
# Ver los datos
X
# Media aritmética
sum(X)/length(X)
#Otra forma de cálculo de media aritmética
mean(X)

También es posible ejecutar varias líneas a la vez. Para ello simplemente se seleccionan las líneas que se desean ejecutar y se pulsa *ctrl-R*. Si desea ejecutar todo el archivo *script*, se selecciona todo con *ctrl-A* y se pulsa *ctrl-R*.

Las instrucciones se pueden ejecutar de forma directa, por ejemplo «mean(X)», o guardar el resultado «Media<-mean(X)». En este segundo caso tendríamos que ejecutar una línea donde escribiésemos «Media» para poder ver el resultado, pero el valor quedaría guardado si quisiéramos usarlo posteriormente.

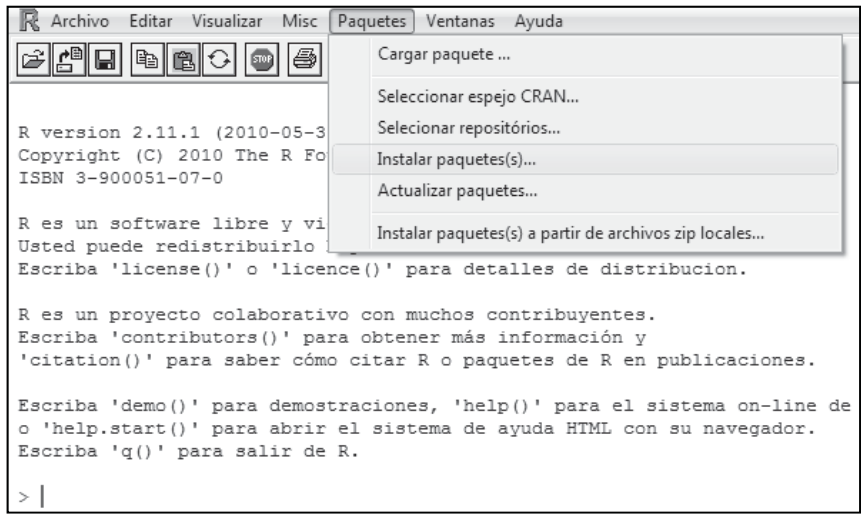
AII.2.3. INSTALAR Y CARGAR PAQUETES Y CONJUNTOS DE DATOS

Para realizar cualquier aplicación matemática con R, existirá un ***paquete,*** o ***librería,*** que es un archivo con el código que contiene las funciones que nos interesa utilizar. Estos paquetes son programados por usuarios que los ponen a disposición pública en los repositorios de «CRAN». Con la instalación básica se incluye una serie de paquetes que nos permitirán realizar numerosas pruebas estadísticas y matemáticas. No obstante, en más de una ocasión nos interesará instalar paquetes adicionales. El modo más elemental de instalar un paquete es ejecutando la siguiente instrucción:

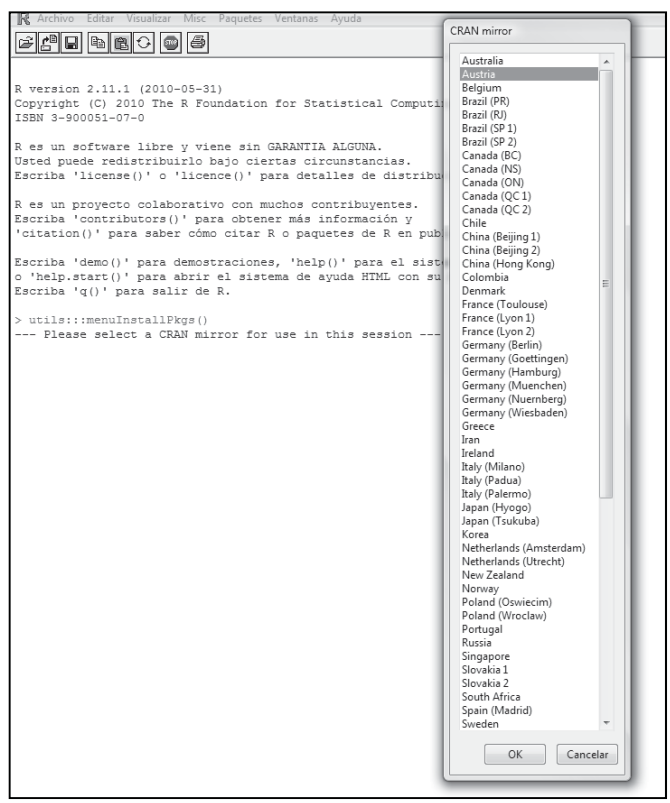
> install.packages("*nombre del paquete*", dep=TRUE)

El nombre del paquete se escribe entre comillas, y después de la coma es muy recomendable incluir el argumento «dep=TRUE», que indica que se instalen también las dependencias de ese paquete. Las dependencias son otros paquetes que contienen elementos necesarios para que funcionen todas las aplicaciones que hay en el paquete que queramos instalar. Sin esas dependencias, algunas aplicaciones de nuestro paquete podrían no funcionar.

Los GUIs suelen incluir algún tipo de menú para instalar paquetes. Por ejemplo, en *RGui* se encuentra en «Paquetes» ⇨ «Instalar paquetes(s)…», y ya instala por defecto todas las dependencias.


Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
Cargar paquete ...
Seleccionar espejo CRAN...
Selecionar repositórios...
Instalar paquetes(s)...
Actualizar paquetes...
Instalar paquetes(s) a partir de archivos zip locales...
R version 2.11.1 (2010-05-3
Copyright (C) 2010 The R Fo
ISBN 3-900051-07-0
R es un software libre y vi
Usted puede redistribuirlo
Escriba 'license()' o 'licence()' para detalles de distribucion.
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
>

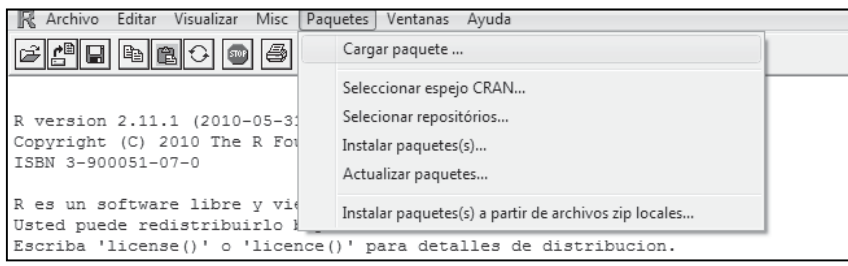
Inmediatamente después de la anterior instrucción, aparece una ventana en la que se nos pide que seleccionemos un servidor "espejo" desde el que podamos descargar el paquete. Suele escogerse por proximidad geográfica. Todos los europeos funcionan bien. Después de seleccionar el servidor, elegimos el paquete en la lista alfabética, y se descargará de forma automática, quedando instalado en ese ordenador de forma permanente.



Una vez instalado el paquete, cuando lo necesitemos es necesario cargarlo. Para ello, se puede escribir alguna de las siguientes instrucciones:

> library(*nombre del paquete*)
> require(*nombre del paquete*)

En los GUIs, también suele haber una opción de menú para cargar paquetes. En el *RGui* de Windows, por ejemplo, sería «Paquetes» ⇨ «Cargar paquete…».



Los ***conjuntos de datos*** consisten en datos reales obtenidos de alguna fuente ya publicada y que están disponibles en *R* para ser manipulados con el objetivo de probar tipos de análisis, gráficos, etc. Muchos de estos conjuntos de datos se encuentran en el paquete "datasets" (cargado por defecto al instalar R y al abrir la sesión) pero también existen en otros. Se cargan mediante la instrucción:

> data(*nombre del conjunto de datos*)

Por ejemplo, cargaremos un conjunto de datos llamado «iris», que se encuentra en "datasets", y que consiste en una serie de datos obtenidos para ejemplares de flores del género *Iris*, en el que se muestran los valores de varios caracteres morfológicos, y la especie a la que pertenece cada ejemplar para el que se efectuaron dichas medidas. Después de cargar el conjunto de datos con la función «data», tecleamos el nombre del conjunto, y se nos muestra su contenido:

```
> data(iris)
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
```

AII.2.4. ACTUALIZACIÓN DE R

Para instalar nuevas versiones de R y no perder los paquetes que se hayan instalado, lo que hay que hacer es, primeramente, desinstalar R. Una vez desinstalado R hay que ir a la carpeta que está en «/Archivos de Programa/R" (en

Windows, o equivalente en otro sistema operativo) y una vez allí, cambiar el nombre de la carpeta por el nombre de la nueva versión disponible en la web. Es muy importante que el nombre coincida exactamente con la nueva versión de R. Por último, instalamos la nueva versión de R. En la instalación el programa preguntará si se desea que R se instale en la carpeta ya existente, a lo cual respondemos «SI». En el caso de que no indique en la instalación que la carpeta ya existe, entonces es que no se cambió el nombre de forma correcta con el nombre de la nueva versión que se instala y, por tanto, no funcionarán los paquetes ya instalados.

También es necesario actualizar los paquetes, para lo cual tendremos que escribir la siguiente instrucción:

```
> update.packages()
```

Entonces R nos muestra una a una las nuevas versiones de paquetes disponibles para actualizar y le diremos que si o que no querremos actualizarlos. También es posible actualizar directamente paquetes concretos, para lo cual habría que modificar la instrucción de la siguiente manera:

```
> update.packages(type="nombre del paquete")
```

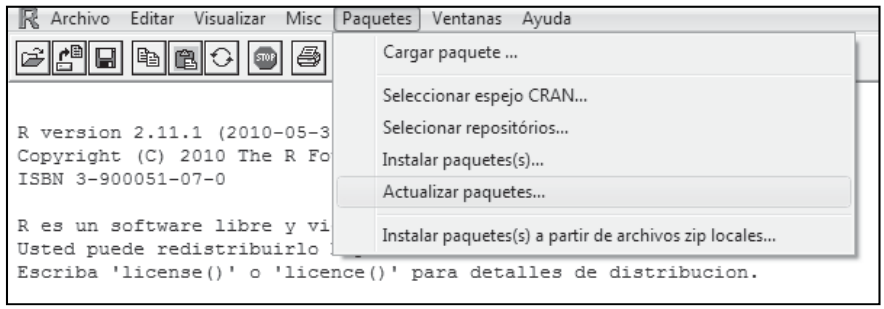
Veamos un ejemplo de actualización de paquetes con instrucciones. Escribimos la instrucción «update.packages()» y tras seleccionar un espejo «CRAN», se nos van mostrando paquetes actualizables, uno por uno, y pulsaremos «y/N/c» si queremos actualizar, no actualizar, o cancelar la actualización en ese momento.

```
> update.packages()
--- Please select a CRAN mirror for use in this session ---
acepack :
 Version 1.3-2.3 installed in C:/ARCHIV~1/R/R-211~1.1/library
 Version 1.3-3.0 available at http://cran.es.r-project.org
Update (y/N/c)?  y
ade4 :
 Version 1.4-14 installed in C:/ARCHIV~1/R/R-211~1.1/library
 Version 1.4-16 available at http://cran.es.r-project.org
Update (y/N/c)?  y
AER :
 Version 1.1-4 installed in C:/ARCHIV~1/R/R-211~1.1/library
 Version 1.1-6 available at http://cran.es.r-project.org
Update (y/N/c)?  y
ape :
 Version 2.5-3 installed in C:/ARCHIV~1/R/R-211~1.1/library
 Version 2.6 available at http://cran.es.r-project.org
Update (y/N/c)?
```

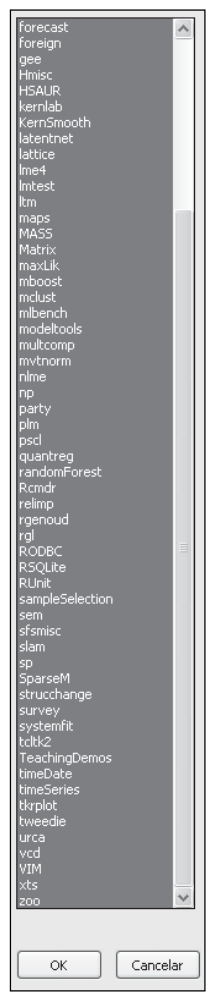
Si actualizamos un paquete específico, con la instrucción que vimos antes, se nos pide en primer lugar seleccionar un espejo «CRAN», y después se actualiza el paquete de forma automática:

```
> update.packages(type="ape")
--- Please select a CRAN mirror for use in this session ---
>
```

Un ejemplo utilizando los menús disponibles en la interfaz RGui. Vamos a «Paquetes» ⇨«Actualizar paquetes…»:



y tras seleccionar un espejo «CRAN» para la sesión, se nos muestra la lista de paquetes instalados, seleccionados en azul aquellos para los cuales existan actualizaciones disponibles:



Pulsaríamos «OK» si queremos actualizar todos, o si no es así, seleccionaríamos solamente aquellos que queremos actualizar y pulsamos «OK».

AII.2.5. AYUDA, EJEMPLOS Y *DEMOS*

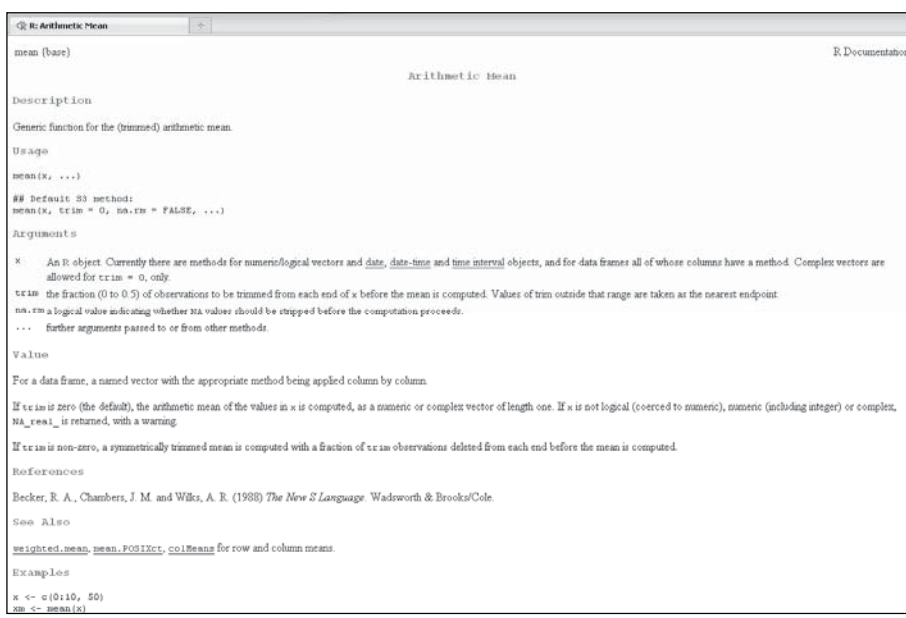
Dentro del programa R, los sistemas para obtener ayuda funcionan muy bien y son imprescindibles para cualquier usuario. En primer lugar disponemos de funciones que nos envían a páginas de ayuda en HTML donde se describen perfectamente las características de lo que buscamos, que tiene que ser una función, un argumento de una función, o conjunto de datos existente en alguno de los paquetes cargados en nuestra sesión. Se puede usar cualquiera de estas dos instrucciones:

> help(*nombre*)
> ?*nombre*

Por ejemplo, queremos saber cosas sobre la función «mean»:

> help(mean)

Se abre una página de ayuda en HTML. Centrado en la zona superior aparece el nombre de esta función, arriba a la izquierda se nos indica entre llaves, en qué paquete se encuentra, y el resto es información muy útil sobre el uso de la función, con algún ejemplo y referencias al final:


R: Arithmetic Mean

mean {base} R Documentation

Arithmetic Mean

Description

Generic function for the (trimmed) arithmetic mean.

Usage

```
mean(x, ...)

## Default S3 method:
mean(x, trim = 0, na.rm = FALSE, ...)
```

Arguments

x	An R object. Currently there are methods for numeric/logical vectors and date, date-time and time interval objects, and for data frames all of whose columns have a method. Complex vectors are allowed for trim = 0, only.
trim	the fraction (0 to 0.5) of observations to be trimmed from each end of x before the mean is computed. Values of trim outside that range are taken as the nearest endpoint.
na.rm	a logical value indicating whether NA values should be stripped before the computation proceeds.
...	further arguments passed to or from other methods.

Value

For a data frame, a named vector with the appropriate method being applied column by column.

If trim is zero (the default), the arithmetic mean of the values in x is computed, as a numeric or complex vector of length one. If x is not logical (coerced to numeric), numeric (including integer) or complex, NA_real_ is returned, with a warning.

If trim is non-zero, a symmetrically trimmed mean is computed with a fraction of trim observations deleted from each end before the mean is computed.

References

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.

See Also

weighted.mean, mean.POSIXct, colMeans for row and column means.

Examples

```
x <- c(0:10, 50)
xm <- mean(x)
```

Además también podemos buscar términos acerca de los que queremos saber qué cosas hay disponibles en R relacionadas con ellos (nótese la diferencia con el caso anterior, en el que no se usan comillas). En este caso la búsqueda se efectúa en todo el material disponible para R, aunque se encuentre en paquetes que no se han cargado ni instalado:

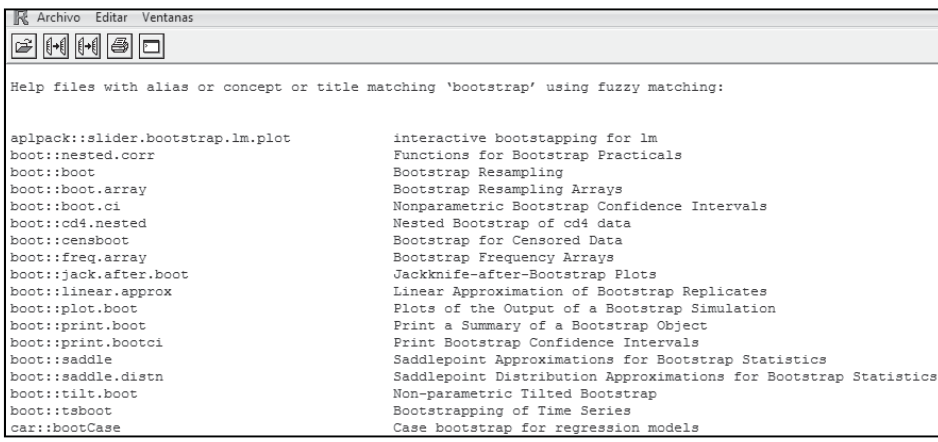
> help.search("*nombre*")

> ??*nombre*

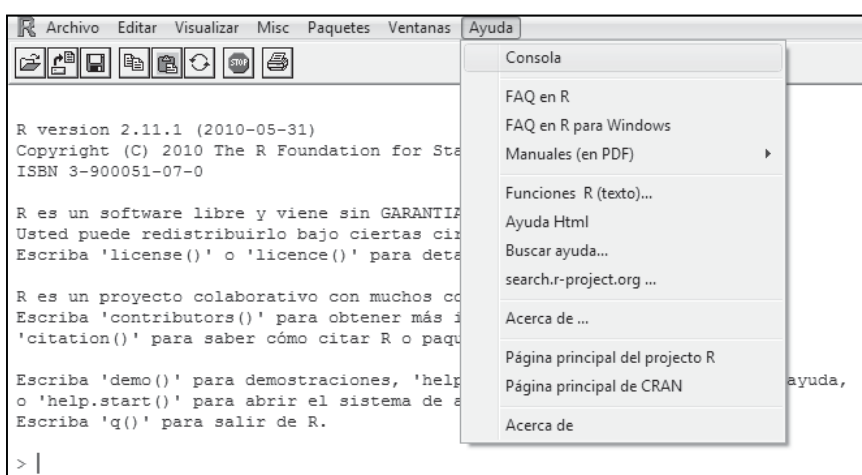
Aparece una ventana con los nombres de todos los términos existentes en R que están relacionados con eso que buscamos, y en qué paquete se encuentran. Si queremos, por ejemplo, buscar información sobre alguna función, o paquetes que sirvan para hacer «bootstrap»:

> help.search("bootstrap")

Se nos muestra el siguiente cuadro con todas las coincidencias que existen en R con ese término. El símbolo «::» indica que la función de la derecha se encuentra en el paquete de la izquierda. Más a la derecha se nos da una breve descripción sobre lo que hace esa función.


Archivo Editar Ventanas
Help files with alias or concept or title matching 'bootstrap' using fuzzy matching:
aplpack::slider.bootstrap.lm.plot interactive bootstapping for lm
boot::nested.corr Functions for Bootstrap Practicals
boot::boot Bootstrap Resampling
boot::boot.array Bootstrap Resampling Arrays
boot::boot.ci Nonparametric Bootstrap Confidence Intervals
boot::cd4.nested Nested Bootstrap of cd4 data
boot::censboot Bootstrap for Censored Data
boot::freq.array Bootstrap Frequency Arrays
boot::jack.after.boot Jackknife-after-Bootstrap Plots
boot::linear.approx Linear Approximation of Bootstrap Replicates
boot::plot.boot Plots of the Output of a Bootstrap Simulation
boot::print.boot Print a Summary of a Bootstrap Object
boot::print.bootci Print Bootstrap Confidence Intervals
boot::saddle Saddlepoint Approximations for Bootstrap Statistics
boot::saddle.distn Saddlepoint Distribution Approximations for Bootstrap Statistics
boot::tilt.boot Non-parametric Tilted Bootstrap
boot::tsboot Bootstrapping of Time Series
car::bootCase Case bootstrap for regression models

Los GUIs, además de tener un menú para conseguir ayuda, poseen a menudo enlaces hacia la página web de R y manuales en pdf donde también se puede encontrar ayuda y FAQs (de este tipo de documentación hablaremos en el siguiente apartado con más detalle). Esto es un ejemplo con RGui:


Archivo Editar Visualizar Misc Paquetes Ventanas Ayuda
Consola
FAQ en R
FAQ en R para Windows
Manuales (en PDF)
Funciones R (texto)...
Ayuda Html
Buscar ayuda...
search.r-project.org ...
Acerca de ...
Página principal del projecto R
Página principal de CRAN
Acerca de
R version 2.11.1 (2010-05-31)
Copyright (C) 2010 The R Foundation for Sta
ISBN 3-900051-07-0
R es un software libre y viene sin GARANTIA
Usted puede redistribuirlo bajo ciertas cir
Escriba 'license()' o 'licence()' para deta
R es un proyecto colaborativo con muchos co
Escriba 'contributors()' para obtener más i
'citation()' para saber cómo citar R o paqu
Escriba 'demo()' para demostraciones, 'help
ayuda,
o 'help.start()' para abrir el sistema de a
Escriba 'q()' para salir de R.
>

También es interesante ver ejemplos de cómo operan determinadas funciones. Esto se haría con la siguiente instrucción:

> example(*nombre de la función*)

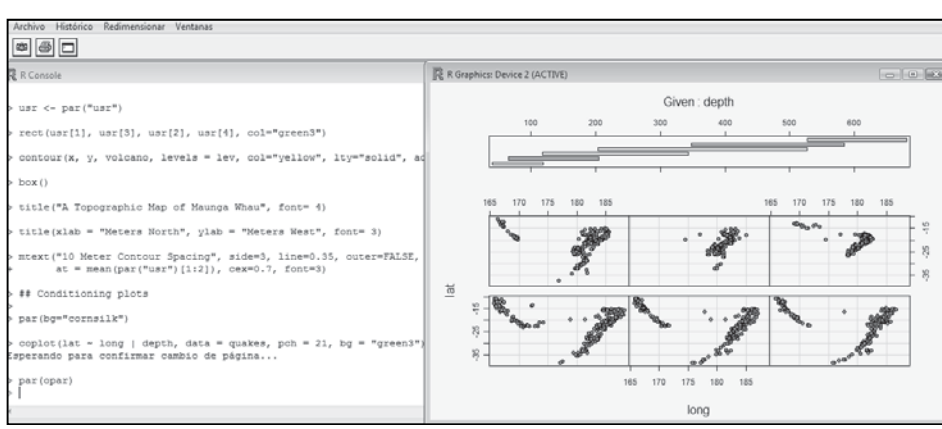
En estos ejemplos se nos muestran conjuntos de datos que son analizados con la función correspondiente, con todo el proceso (paquetes a instalar, ecuaciones, obtención de resultados, gráficos, etc.). He aquí un ejemplo con la función del modelo lineal generalizado:

```
> example(glm)

glm> ## Dobson (1990) Page 93: Randomized Controlled Trial :
glm> counts <- c(18,17,15,20,10,20,25,13,12)

glm> outcome <- gl(3,1,9)

glm> treatment <- gl(3,3)

glm> print(d.AD <- data.frame(treatment, outcome, counts))
  treatment outcome counts
1         1       1     18
2         1       2     17
3         1       3     15
4         2       1     20
5         2       2     10
6         2       3     20
7         3       1     25
8         3       2     13
9         3       3     12

glm> glm.D93 <- glm(counts ~ outcome + treatment, family=poisson())

glm> anova(glm.D93)
Analysis of Deviance Table
```

Los *demo* muestran un conjunto de cosas que sería posible hacer con R en un aspecto determinado (gráficos, análisis de datos de algún tipo concreto, etc.) Por ejemplo, si tecleamos:

> demo(graphics)

Este *demo* nos muestra diversos tipos de gráficos que es posible realizar, y en la consola aparecen las instrucciones necesarias para generar los gráficos que vemos desplegados a medida que pulsamos con el ratón.


Archivo Histórico Redimensionar Ventanas
R Console
> usr <- par("usr")
> rect(usr[1], usr[3], usr[2], usr[4], col="green3")
> contour(x, y, volcano, levels = lev, col="yellow", lty="solid", ac
> box()
> title("A Topographic Map of Maunga Whau", font= 4)
> title(xlab = "Meters North", ylab = "Meters West", font= 3)
> mtext("10 Meter Contour Spacing", side=3, line=0.35, outer=FALSE,
+ at = mean(par("usr")[1:2]), cex=0.7, font=3)
> ## Conditioning plots
> par(bg="cornsilk")
> coplot(lat ~ long | depth, data = quakes, pch = 21, bg = "green3")
Esperando para confirmar cambio de página...
> par(opar)
R Graphics: Device 2 (ACTIVE)
Given : depth
lat
long

AII.2.6. DOCUMENTACIÓN SOBRE R

En la propia página web del proyecto R existen enlaces a una serie de manuales gratuitos en pdf. En el margen izquierdo, donde dice *Documentation*, podemos seguir el enlace a *Manuals* donde encontraremos el pdf de diversos manuales en inglés, editados por el *R Development Core Team*. Debajo de *Manuals*, tenemos una sección de preguntas frecuentes *(FAQs*) y debajo de estas, el *R journal* con noticias y novedades sobre *R.* Debajo también vemos el enlace a una "wiki" sobre R. Más abajo, en *Books* se hace una reseña de diversos libros publicados sobre R, que no son gratuitos. Más abajo, es interesante *Other*, donde se muestran enlaces hacia otra serie de documentos, en inglés y otras lenguas, de cuyo contenido no se responsabiliza el *R Development Core Team*, pero que se consideran interesantes.

En el apartado de *Software*, en *Packages*, tenemos una lista con todos los paquetes disponibles. Si seguimos el enlace a cada uno de ellos se nos muestra información entre la cual se encuentra un manual en pdf que explica el funcionamiento del paquete.

AII.3. R Commander

R commander es una Interfaz Gráfica de Usuario (GUI) que facilita considerablemente el uso de R. Contiene un menú que permite –de forma similar a otros programas estadísticos– realizar de modo sencillo la mayoría de las operaciones estadísticas habituales, y separa las instrucciones y resultados en dos ventanas diferentes.

AII.3.1. CARACTERÍSTICAS DE R COMMANDER

Para utilizar R Commander hay que instalar el paquete "Rcmrd" (véase apartado AII.2.3). En general al mismo tiempo se instalan automáticamente todos los paquetes dependientes o complementarios necesarios para un funcionamiento correcto. Los diseñadores de R Commander trataron de incluir en él todas las técnicas contenidas en un curso básico de Estadística, añadiendo posteriormente algunas otras técnicas de uso común, pero después han preferido mantener su estructura simple en lugar de seguir añadiendo técnicas en un menú crecientemente complejo que intente cubrir todos los métodos.

Posteriormente han surgido nuevos paquetes que se incorporan como opciones de menú complementarias o extensiones de R Commander, de gran utilidad para usuarios que reclaman técnicas específicas en el menú, los cuales solo deben instalar el paquete adicional (*plug-in package*) de su interés, como por ejemplo "FactoMineR", utilizado en el capítulo XIII de este libro. La mayoría de los paquetes *plug-in* de R Commander empiezan por "RcmdrPlugin.", como por ejemplo "RcmdrPlugin.MAd" para Meta análisis, "RcmdrPlugin.epack" para series temporales, "RcmdrPlugin.qual" para control de calidad, "RcmdrPlugin.survival" para análisis de supervivencia, o "RcmdrPlugin.DoE" para diseño de experimentos industriales.

Las opciones de menú de R Commander conducen en general a cuadros de diálogo que permiten incorporar a las instrucciones todas las variables y opciones necesarias en cada aplicación, de manera que el usuario de R no necesita conocer ni utilizar el lenguaje de programación. La estructura de los cuadros de diálogo es simple y clara, y a menudo no necesita explicación alguna. Habitual-

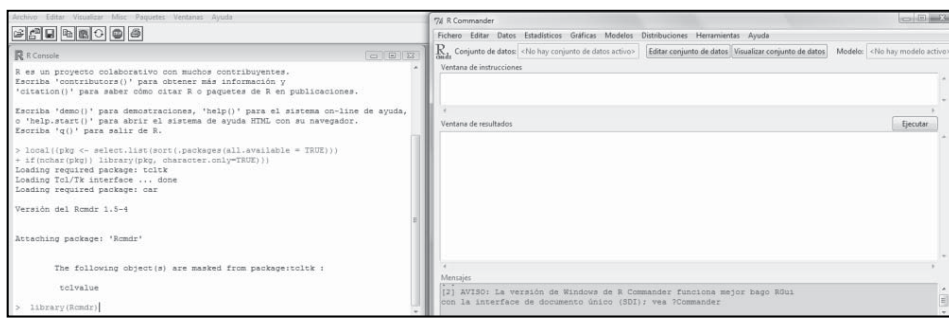
mente incluye un botón de ayuda que remite a la documentación de R sobre la función que se quiere utilizar.

La principal ventaja de R Commander es que el usuario puede realizar casi todas las operaciones estadísticas de tratamiento de datos sin conocer en absoluto el complejo lenguaje de programación. R Commander se encarga de construir las órdenes a partir del menú, como lo hacen otros programas estadísticos (SPSS, Statística, SAS, etc.). Incluso para aquellos que conocen a fondo el lenguaje de R resulta a menudo más cómodo y rápido utilizar R Commander.

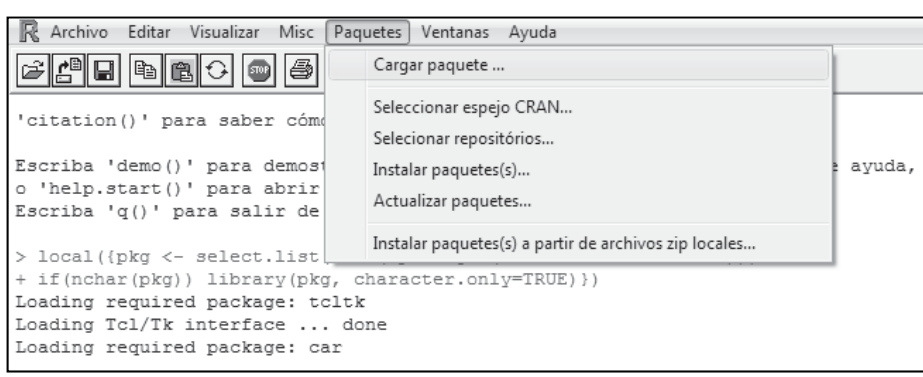
Las limitaciones de R Commander consisten en que muchas técnicas estadísticas no están en el menú (solo las más habituales), y que algunas opciones de las técnicas incluidas no se pueden aplicar desde el menú. El conocimiento del lenguaje de programación de R permite multiplicar la potencia estadística y mejorar los resultados y su presentación, para lo cual se pueden modificar convenientemente las órdenes generadas inicialmente por R Commander.

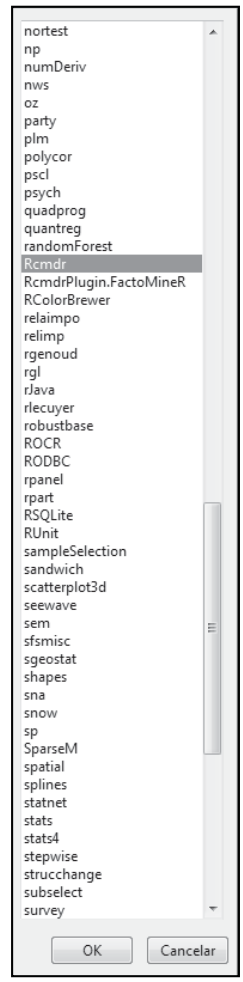
Para iniciar R Commander tecleamos –recuerde que en R siempre es necesario respetar las mayúsculas– la instrucción siguiente en la consola de R (RGui):

```
> library(Rcmdr)
```


R Console
R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.
> local({pkg <- select.list(sort(.packages(all.available = TRUE)))
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
Loading required package: tcltk
Loading Tcl/Tk interface ... done
Loading required package: car
Versión del Rcmdr 1.5-4
Attaching package: 'Rcmdr'
The following object(s) are masked from package:tcltk :
tclvalue
> library(Rcmdr)
R Commander
Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda
Conjunto de datos: <No hay conjunto de datos activo>
Editar conjunto de datos
Visualizar conjunto de datos
Modelo: <No hay modelo activo>
Ventana de instrucciones
Ventana de resultados
Ejecutar
Mensajes
[2] AVISO: La versión de Windows de R Commander funciona mejor bajo RGui
con la interface de documento único (SDI); vea ?Commander

También es posible iniciar R Commander en el menú «Paquetes» ⇨ «Cargar paquete…», y en la ventana que aparece se selecciona «Rcmdr».





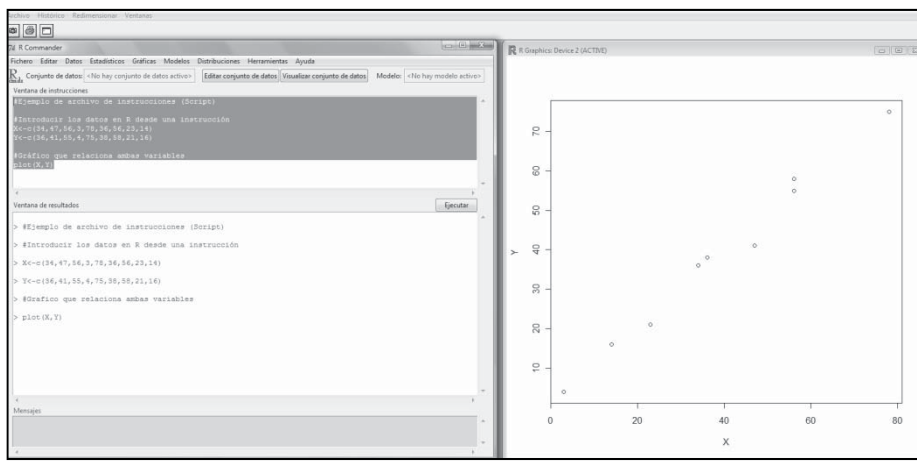
Cuando se utiliza el menú de R Commander, el programa genera las instrucciones adecuadas, que se muestran en la «Ventana de instrucciones». También aparecen las instrucciones junto con los resultados producidos en la «Ventana de resultados». El usuario puede acceder a cualquiera de ellas para observar, copiar resultados o instrucciones, o modificar y ejecutar de nuevo cualquier instrucción (con el botón «Ejecutar»), de un modo mucho más interactivo que con el RGui. Es más, cualquier instrucción o conjunto de instrucciones que podamos construir en RGui puede ser también generalmente realizada en la ventana de instrucciones de R Commander.

La separación de instrucciones y resultados en ventanas diferentes es muy conveniente para una mejor organización de todos los elementos de la aplicación estadística. Además R Commander muestra una tercera ventana de «Mensajes», en la que advierte de cualquier error o problema en la aplicación. Los mensajes se numeran correlativamente, y pueden ser borrados en cualquier momento para evitar confusión e identificar más claramente los mensajes nuevos.

Veamos el siguiente ejemplo, cuyas instrucciones las introducimos en la «Ventana de instrucciones». Después de seleccionarlo todo y pulsar Ejecutar

aparecen las instrucciones junto con los resultados producidos en la «Ventana de resultados». El gráfico aparece en la ventana principal de R.

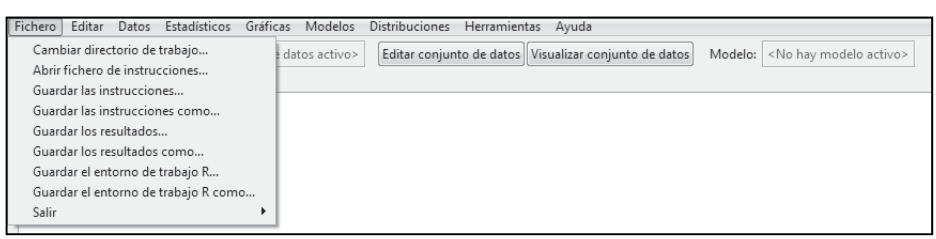
```
#Ejemplo de archivo de instrucciones (Script)
#Introducir los datos en R desde una instrucción
X<-c(34,47,56,3,78,36,56,23,14)
Y<-c(36,41,55,4,75,38,58,21,16)
#Gráfico que relaciona ambas variables
plot(X,Y)
```



AII.3.2. EL MENÚ DE R COMMANDER

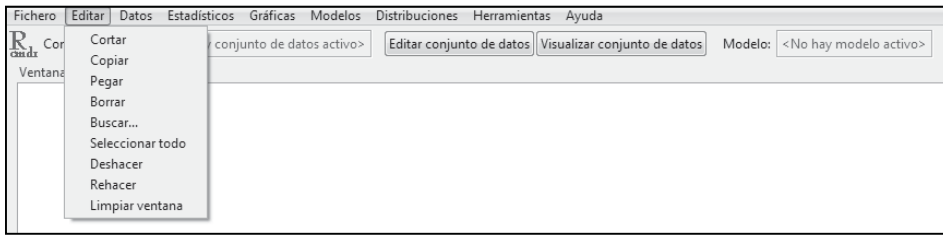
AII.3.2.1. Fichero

Permite «Cambiar» (establecer) el directorio o carpeta de trabajo, «Abrir» un archivo de instrucciones, «Guardar las instrucciones» en diversos formatos, «Guardar los resultados», «Guardar el entorno de trabajo R» y «Salir» de R Commander.



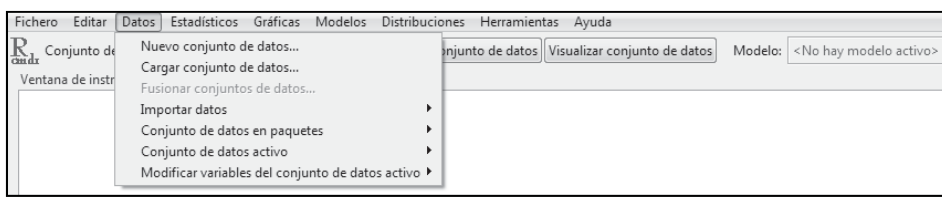
AII.3.2.2. Editar

Permite las operaciones habituales de esta opción de menú en cualquier aplicación Windows: «Cortar», «Copiar», «Pegar», «Borrar», «Buscar», «Seleccionar todo», «Deshacer/Rehacer», o «Limpiar ventana». Esta última vacía la ventana de instrucciones o la de resultados, aquella en la que esté situado el cursor.

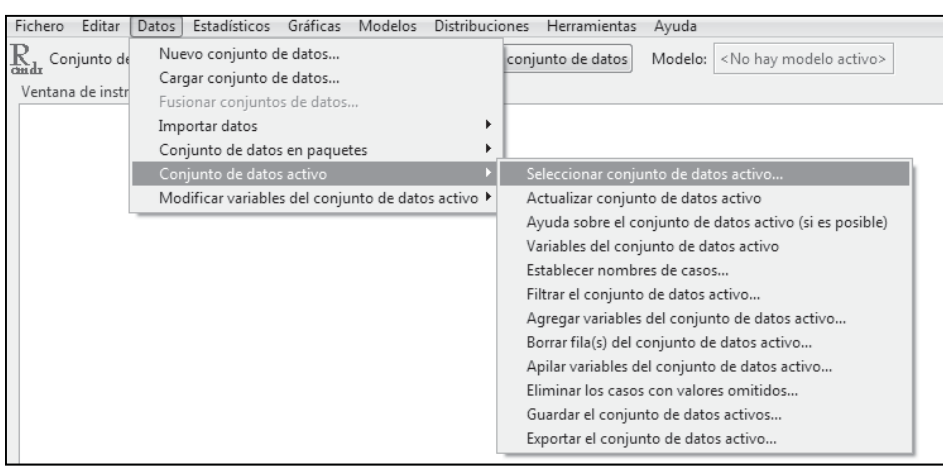


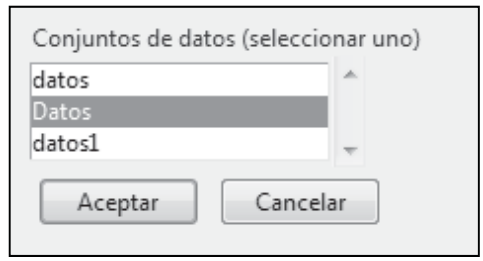
AII.3.2.3. Datos

«Nuevo conjunto de datos», sirve para introducir un conjunto de datos directamente (esto es útil cuando el conjunto de datos es pequeño); también se puede «Cargar un conjunto de datos» en formato R; «Fusionar conjuntos de datos»; «Importar datos» desde archivo de texto o portapapeles, o desde otros programas estadísticos (SPSS, Minitab, STATA….), o desde Excel, Access o dBase.



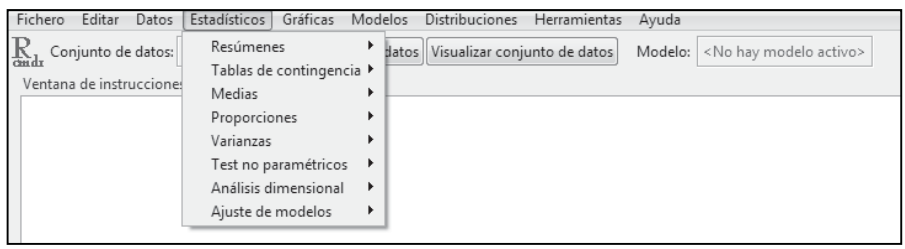
También se puede seleccionar el «Conjunto de datos activo», o «Modificar las variables del conjunto de datos activo» (recodificar, calcular una nueva variable, tipificar, convertir, segmentar, reordenar, renombrar, eliminar, ...). Las instrucciones se ejecutan sobre el conjunto de datos activo. Cuando existen varios, lo cual es muy habitual, debe definirse cuál es el activo para que el menú de R Commander se refiera siempre a él. Por ejemplo, supongamos que tenemos dos series de datos, vamos a «Datos» ⇨ «Conjunto de datos activo» ⇨ «Seleccionar conjunto de datos activo…», y en la ventana que aparece se selecciona la serie de datos con la que se quiera trabajar.





AII.3.2.4. Estadísticos

Es sin duda la opción principal del menú de R Commander, ya que todas las operaciones estadísticas de tratamiento de datos están aquí.



En «Resúmenes» se calculan medidas estadísticas, tablas de frecuencias, coeficientes de correlación, y realiza algunos contrastes sencillos (Test de correlación, Test de Normalidad de Shapiro-Wilk).

En «Tablas de contingencia» se construyen tanto tablas de doble entrada como de entradas múltiples, incluyendo algunos elementos de análisis (contraste Chi cuadrado de independencia, cálculo de *Odds Ratio* en tablas 2x2).

En «Medias» se realizan las pruebas *t* para una o dos muestras, así como el ANOVA de uno o varios factores.

En la opción «Proporciones» se pueden contrastar proporciones de una o dos muestras.

En «Varianzas» se puede realizar el test *F* para comparar varianzas, así como el test de homogeneidad de Bartlett o el test de Levene.

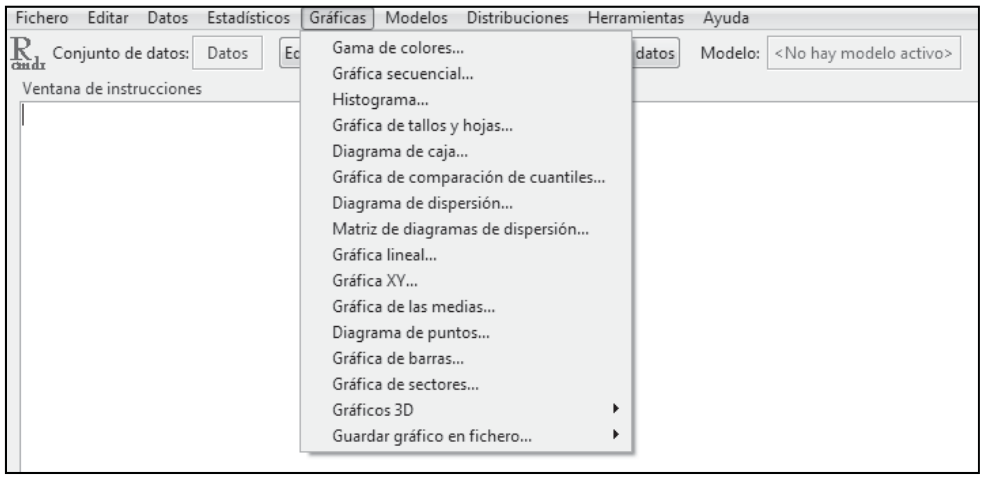
En «Test no paramétricos» están el Test de Wilcoxon para dos muestras independientes o emparejadas, el test de Kruskal-Wallis (similar al ANOVA de un factor con más de dos niveles) y el test de Friedman.

La opción «Análisis dimensional» agrupa varias técnicas estadísticas multivariantes: Fiabilidad de escala, Análisis de Componentes Principales, Análisis Factorial, Análisis de Agrupación por k-medias, Agrupación jerárquica.

Por último en «Ajuste de modelos» encontramos cálculo de regresiones, construcción de modelos lineales, modelo logit multinomial, o modelo ordinal de regresión.

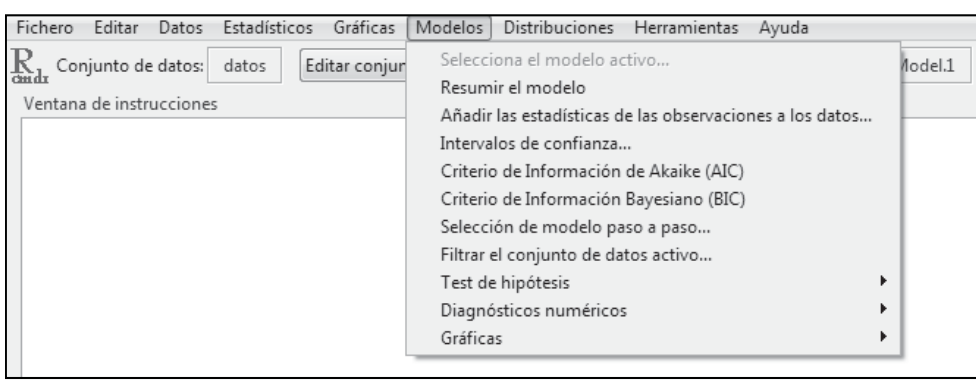
AII.3.2.5. Gráficas

Contiene diferentes gráficos: secuencial (para series temporales), histograma (para variables continuas), de tallo y hojas, diagrama de cajas, diagrama de dispersión (para ver la relación entre dos variables cuantitativas), gráfica de líneas, de barras, de sectores (para variables cualitativas), gráficos 3D, etc.



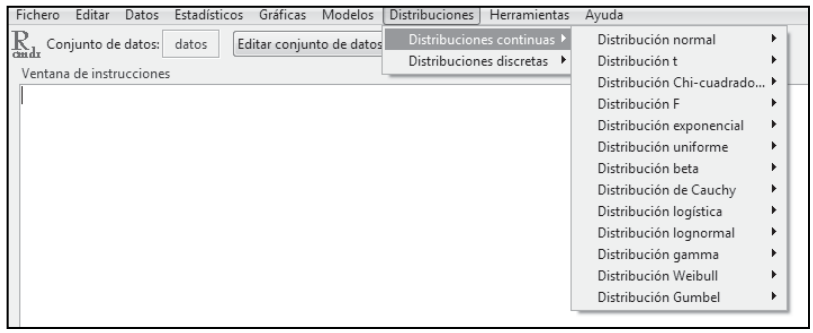
AII.3.2.6. Modelos

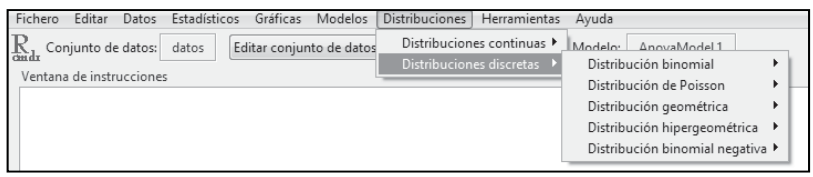
Permite realizar múltiples operaciones relativas a un modelo construido previamente, como elaborar una tabla ANOVA, comparar varios modelos y elegir el más adecuado utilizando los criterios de información (AIC, BIC), aplicar modelos paso a paso (por ejemplo en el Análisis Discriminante), construir contrastes, y realizar diagnósticos (FIV, Test de Breusch-Pagan para heteroscedasticidad, Test de Durbin-Watson para autocorrelación, diagnósticos gráficos de residuos). Para que aparezcan todas las opciones del menú es necesario que anteriormente se realizara algún modelo, por ejemplo haber realizado un ANOVA.


Fichero Editar Datos Estadísticos Gráficas Modelos Distribuciones Herramientas Ayuda
Conjunto de datos: datos
Editar conjur
Ventana de instrucciones
Selecciona el modelo activo...
Resumir el modelo
Añadir las estadísticas de las observaciones a los datos...
Intervalos de confianza...
Criterio de Información de Akaike (AIC)
Criterio de Información Bayesiano (BIC)
Selección de modelo paso a paso...
Filtrar el conjunto de datos activo...
Test de hipótesis
Diagnósticos numéricos
Gráficas

AII.3.2.7. Distribuciones

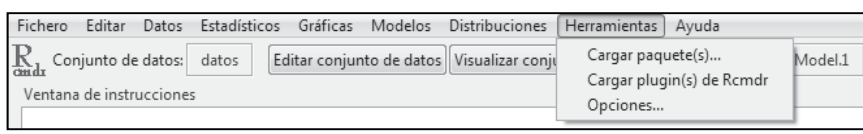
Se utiliza para calcular probabilidades o cuantiles, realizar gráficos, o simular muestras de diferentes distribuciones continuas (Normal, *t*, Chi-cuadrado, *F*, exponencial, uniforme, beta, Cauchy, logística, lognormal, gamma, Weibull, Gumbel) o discretas (binomial, Poisson, geométrica, hipergeométrica, binomial negativa).





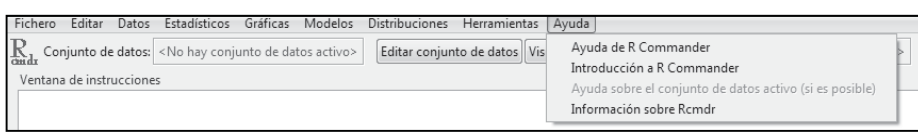
AII.3.2.8. Herramientas

Sirve para cargar paquetes, cargar *plug-ins* de R Commander, y para modificar algunas opciones de configuración del propio R Commander. Los paquetes deben haber sido previamente instalados desde la consola de R o RGui; cuando se carga un *plug-in*, la próxima vez –y sucesivas– que se inicie R Commander aparecerá incorporado al menú.



AII.3.2.9. Ayuda

Este último menú, permite encontrar información sobre términos o funciones de R, tiene un pequeño manual de Introducción a R Commander, facilita ayuda sobre el conjunto de datos activo (si tenemos uno), y sobre el propio paquete Rcmdr.



AII.4. El lenguaje R

AII.4.1. APUNTES PRELIMINARES

R es un lenguaje de programación orientado a objetos, a los cuales se les asigna una serie de características. Antes de empezar, es necesario resaltar al-

gunas reglas básicas para escribir instrucciones, de las que mostraremos ejemplos más abajo:

- Se distingue entre mayúsculas y minúsculas.
- El símbolo «>» aparece antes de escribir cualquier instrucción, indicando que todo está correcto para recibirla. Si lo que se muestra es un resultado, no aparece este símbolo.
- La tecla *Enter* se pulsa siempre para cambiar de línea y nos muestra los resultados de la instrucción anterior, si es el caso.
- El símbolo «#» anula la ejecución de lo que viene después, de modo que se utiliza antepuesto a comentarios.
- Si un comando es correcto sintácticamente pero falta algo para terminarlo (típicamente un paréntesis), aparece un símbolo «+» en la siguiente línea, tras el cual se puede escribir lo que deseemos.
- Si escribimos diferentes instrucciones en una misma línea, se separan con «;».
- La asignación puede hacerse con «<-», «<<-» o «=», pero el signo «=» también se usa con otras funciones, por lo que es recomendable no utilizarlo para este fin.
- Si escribimos instrucciones en distintas líneas, que queramos agrupar, se agrupan con llaves «{}».
- La flecha hacia arriba en el teclado nos permite reescribir instrucciones de líneas superiores.

Por ejemplo, queremos crear un objeto que llamaremos «x» y al que le queremos asignar el valor de la expresión matemática «2+5», teclearíamos:

```
> x<-2+5
```

Las características asignadas a «x» quedarían almacenadas en la memoria, y si tecleamos «x» y pulsamos *Enter*, se nos devuelve su valor:

```
> x
```

```
> x <- 2+ 5
> x
[1] 7
```

El «[1]» indica que como resultado de «x», se nos devuelve un único valor. Si tecleásemos «X» mayúscula, se nos diría que tal objeto no existe:

```
> X
Error: objeto 'X' no encontrado
```

El símbolo «#» sirve para escribir un comentario, por ejemplo, qué es el objeto «x», antes de asignarle un valor:

```
> # x es el resultado de la suma de dos números
> x<-2+5
> x
```

```
> # x es el resultado de la suma de dos números
> x<-2+5
> x
[1] 7
```

Creando otro objeto, que llamaremos «y», nos hemos olvidado algo, en concreto cerrar un paréntesis, aparecerá entonces un símbolo «+», y si cerramos el paréntesis en esa línea, todo permanece correcto:

```
> y <- x+(3+8
+ )
> y
```

```
> y <- x+(3+8
+ )
> y
[1] 18
```

Si queremos escribir dos instrucciones en una misma línea, por ejemplo, para definir dos objetos nuevos «z» y «w» utilizamos «;». En este ejemplo se considera que los objetos «x» e «y» están almacenados en la memoria, es decir, seguiríamos en la misma sesión en la que dichos objetos se crearon, o en la misma área de trabajo. Vemos que tecleando «z» y «w» se nos devuelve su valor:

```
> z <- 2+9; w <- x+y+z
> z;w
```

```
> z <- 2+9; w <- x+y+z
> z;w
[1] 11
[1] 36
```

Para agrupar instrucciones en varias líneas, utilizaríamos llaves, por ejemplo, porque tenemos una expresión muy larga que queremos fragmentar en componentes diferenciados. Esto es algo habitual escribiendo funciones. Aunque todavía no hemos mostrado cómo escribir funciones, he aquí un ejemplo de una función sencilla.

```
> mifuncion<-function(x){
+ y<-sum(1:5)^3
+ z<-log(x)
+producto<-y*z
+return(producto)
+ }
> mifuncion
```

```
> mifuncion<-function(x){
+ y<-sum(1:5)^3
+ z<-log(x)
+ producto<-y*z
+ return(producto)
+ }
> mifuncion(45)
[1] 12847.49
```

AII.4.2. IMPORTACIÓN DE DATOS

Con mucha frecuencia, en R trabajaremos con archivos de datos importados de otras fuentes. Existen numerosos formatos desde los que es posible importar datos: txt, csv, xls, SPSS, ODBC, etc. Para algunas de ellas es necesario instalar paquetes específicos. Iremos viendo las funciones más habituales.

Para leer datos en formato tabla (por ejemplo archivos txt) se usa la función «read.table()». Esta función asume que las columnas corresponden a cada variable y las filas son los distintos casos. Esta función estaría importando automáticamente un *data frame*. Un argumento importante es «header», al que podemos definir como verdadero o falso («TRUE» o «FALSE»). Este argumento nos indica si nuestro archivo contiene o no, en el encabezado de cada columna, el nombre de la variable. Otro argumento muy importante es «sep», que indica el separador de campos. Por defecto, el separador es el espacio en blanco. Esto crea un problema si existen valores perdidos, por lo que sería necesario utilizar otro separador, como por ejemplo la coma «,». Por último, la instrucción «encoding="latin1"» permite importar nombres con acentos.

>read.table("*Nombre del archivo.extensión*",header=…...,sep=……., encoding= "latin1")

Vamos a ver un ejemplo en el cual queremos importar un archivo **Lagunas.txt** con datos fisicoquímicos de diferentes lagunas, que se encuentra en la carpeta Apéndice del CD. Este archivo de texto se guardó a partir del archivo de Excel **Lagunas.xlsx** en el formato texto delimitado por tabulaciones. Tenemos encabezado, y no modificamos el separador de campos. Debido a que en el archivo hay celdas vacías, es necesario incluir la instrucción «fill=TRUE» para que lea el archivo de forma correcta. Aparecen los nombres de las variables como encabezado y las filas numeradas como casos.

> datos<-read.table("Lagunas.txt",header=TRUE,fill=TRUE,encoding= "latin1")
> datos

```
> datos<-read.table("Lagunas.txt",header=TRUE,fill=TRUE,encoding= "latin1")
> datos
   Laguna Estación Temperatura Conductividad    pH Nitrato Nitrito  Amonio Fosfato Silicato
1       1        1        29,1           495  6,08       0    1,09    1,14    0,32    62,78
2       2        1        28,1          1537  7,85    0,52    0,28    0,94   0,384  231,673
3       2        2        28,6          1568 0,253   0,839   0,428 262,306
4       3        1        28,2           755  8,52    0,52   0,275   0,937   0,384  231,673
5       3        2        28,5           739  8,24       0   0,348   0,339   0,761  172,686
6       3        3        28,6           748   8,6    0,13   0,386   0,419   0,867  189,409
7       4        1        27,5           905  7,63       0   0,893   0,915   1,056   91,699
8       5        1        26,4           419  7,72   0,708   0,668   0,258 113,179
9       6        1        23,9          1034  7,64    0,42   0,301   1,229   0,263  189,038
10      7        1        32,6           217  7,42     0,3   0,843   2,076   0,618  117,996
11      8        1        31,3           371  7,32       0   0,697   0,635   0,489   99,225
```

El problema es que cuando existen celdas vacías, como en este caso, a la hora de mostrar los resultados se observa que no deja el hueco de las celdas sin datos y, por lo tanto, el archivo no ha sido leído correctamente.

Veamos ahora el caso en el que el mismo archivo está guardado en formato csv (**Lagunas.csv**), el cual se grabó en el formato delimitado por comas, es decir, los decimales son comas. Cuando el separador decimal es la coma en lugar del punto, como en este caso, se utiliza como delimitador de campos el punto y coma.

```
> datos<-read.table("Lagunas.csv",header=TRUE,sep=";",encoding="latin1")
> datos
```

```
> datos<-read.table("Lagunas.csv",header=TRUE,sep=";",encoding="latin1")
> datos
  Laguna Estación Temperatura Conductividad   pH Nitrato Nitrito Amonio Fosfato Silicato
1      1        1        29,1           495 6,08       0    1,09   1,14    0,32    62,78
2      2        1        28,1          1537 7,85    0,52    0,28   0,94   0,384  231,673
3      2        2        28,6          1568                0,253  0,839   0,428  262,306
4      3        1        28,2           755 8,52    0,52   0,275  0,937   0,384  231,673
5      3        2        28,5           739 8,24       0   0,348  0,339   0,761  172,686
6      3        3        28,6           748  8,6    0,13   0,386  0,419   0,867  189,409
```

Se observa que no es necesario incluir la instrucción «fill=TRUE» para que reconozca las celdas que están vacías y, además, deja el espacio de las celdas que no tienen datos. Esta opción de importar datos es posiblemente la mejor cuando la configuración del ordenador sea en español, aunque también es muy útil la importación con R Commander que veremos más adelante.

Otra función muy útil, que se utiliza exclusivamente para importar datos en formato csv es «read.csv()». Para esta función, el separador de campos por defecto es un espacio, al igual que para «read.table()». No obstante, existe la variante «read.csv2()» cuando el separador de decimales es la coma y el de campos el punto y coma.

```
>read.csv("Nombre del archivo.csv",header=… ,sep=… )
>read.csv2("Nombre del archivo.csv",header=… ,sep=… )
```

Para el archivo **Lagunas.csv**, necesitamos la segunda función «read.csv2()», ya que el separador es punto y coma, y con esta función ese es el separador que se admite por defecto.

```
>datos<-read.csv2("Lagunas.csv",header=TRUE,encoding="latin1")
> datos
```

```
> datos<-read.csv2("Lagunas.csv",header=TRUE,encoding="latin1")
> datos
  Laguna Estación Temperatura Conductividad   pH Nitrato Nitrito Amonio Fosfato Silicato
1      1        1        29.1           495 6.08    0.00   1.090  1.140   0.320   62.780
2      2        1        28.1          1537 7.85    0.52   0.280  0.940   0.384  231.673
3      2        2        28.6          1568   NA      NA   0.253  0.839   0.428  262.306
4      3        1        28.2           755 8.52    0.52   0.275  0.937   0.384  231.673
5      3        2        28.5           739 8.24    0.00   0.348  0.339   0.761  172.686
6      3        3        28.6           748 8.60    0.13   0.386  0.419   0.867  189.409
7      4        1        27.5           905 7.63    0.00   0.893  0.915   1.056   91.699
8      5        1        26.4           419 7.72      NA   0.708  0.668   0.258  113.179
```

Se observa que no es necesario incluir la instrucción «fill=TRUE» para que reconozca las celdas que están vacías, pero las identifica como «NA». Esta forma

de importar datos es muy útil para archivos con configuración en español, aunque las celdas vacías las identifica como «NA» (valor omitido o faltante).

Otra función interesante es la que nos permite leer hojas de cálculo de Excel, con una instrucción muy similar a las anteriores. Se trata de «read.xls()». Pero para utilizarla, es necesario instalar y cargar el paquete "xlsReadWrite", y en la mayoría de los casos también aplicar la instrucción «xls.getshlib()». Los argumentos de esta función son similares aunque tienen diferentes nombres. En lugar de «header», tenemos el argumento «colnames», que tiene el mismo significado, pero con la ventaja de que por defecto admite nombres de variables en el encabezado de las columnas. Por defecto, los datos se toman de la hoja 1 del libro de Excel. Si nos interesase leer datos de otra hoja, sería necesario especificar el número de hoja en el argumento «sheet». El problema que tiene esta instrucción que importa archivos de Excel es que, si en la ruta de acceso existen carpetas o archivos con acentos, no encuentra bien el archivo y da error. Esto a veces se soluciona especificando la ruta completa donde está el archivo, aunque a veces incluso así da error. Por tanto, si tuviésemos los datos anteriores guardados como libro de Excel, las instrucciones serían las siguientes, suponiendo que el archivo está en la unidad C:/ del disco duro:

```
> library(xlsReadWrite)
> xls.getshlib()
> datos<-read.xls("C:/Lagunas.xls")
> datos
```

```
> datos<-read.xls("C:/Lagunas.xls")
> datos
  Laguna Estación Temperatura Conductividad   pH Nitrato Nitrito Amonio Fosfato Silicato
1      1        1        29.1           495 6.08    0.00   1.090  1.140   0.320   62.780
2      2        1        28.1          1537 7.85    0.52   0.280  0.940   0.384  231.673
3      2        2        28.6          1568   NA      NA   0.253  0.839   0.428  262.306
4      3        1        28.2           755 8.52    0.52   0.275  0.937   0.384  231.673
5      3        2        28.5           739 8.24    0.00   0.348  0.339   0.761  172.686
6      3        3        28.6           748 8.60    0.13   0.386  0.419   0.867  189.409
```

También es posible leer hojas de cálculo de Excel en formato xlsx (Office 2007). La forma más directa es con la función «read.xlsx()» del paquete "xlsx", que hay que tener instalado. También es necesario tener instalado el programa Java para que la instrucción funcione. Es necesario mencionar que puede dar errores si hay celdas vacías y no reconoce bien los acentos. La función «read.xlsx()», además del nombre del archivo, requiere como argumento el número de la hoja del archivo xlsx que se desea leer. Utilizamos el archivo del ejemplo anterior (en el que no hay celdas vacías), guardado con la extensión xlsx, e indicando que queremos leer la hoja 1:

```
> library(xlsx)
> datos<-read.xlsx("Lagunas.xlsx",1)
> datos
```

```
> datos<-read.xlsx("Lagunas.xlsx",1)
> datos
  Laguna EstaciÃ³n Temperatura Conductividad   pH Nitrato Nitrito Amonio Fosfato Silicato
1      1         1        29.1           495 6.08    0.00   1.090  1.140   0.320   62.780
2      2         1        28.1          1537 7.85    0.52   0.280  0.940   0.384  231.673
3      2         2        28.6          1568 7.50    0.00   0.253  0.839   0.428  262.306
4      3         1        28.2           755 8.52    0.52   0.275  0.937   0.384  231.673
5      3         2        28.5           739 8.24    0.00   0.348  0.339   0.761  172.686
```

Existen más posibilidades para leer otros formatos tipo hoja de cálculo. Un grupo interesante de funciones se encuentra en el paquete "foreign". Entre ellas, existe una para leer hojas de SPSS. Veamos un ejemplo. En primer lugar es necesario instalar y cargar el paquete "foreign", de la manera que se explicó en el apartado AII.2. Nuestro archivo de datos contiene los resultados de un análisis de toxinas presentes en una especie de microalga durante varios días:

```
> library(foreign)
> datos<-read.spss("Cuadro IX.1.sav")
> datos
```

```
> datos<-read.spss("Cuadro IX.1.sav")
> datos
$Especie
  [1] A. minutum   A. minutum   A. minutum   A. minutum
  [8] A. minutum   A. minutum   A. minutum   A. minutum
```

Entre los argumentos de esta función (que como sabemos, siempre pueden consultarse en la página de ayuda) son especialmente interesantes dos: «use.value.labels», que por defecto es «TRUE», es decir, que se usan las etiquetas de valores por defecto, y otro es «to.data.frame», por defecto «FALSE» y, por tanto, no se importa como un *data frame*, sino como una lista.

Dentro del mismo paquete, "foreign", existen más funciones para importar datos desde otros formatos menos habituales como Minitab, Stata, etc. Existen, además, otros paquetes como el "RODBC", que nos permiten importar también datos de cualquier formato de tipo ODBC.

AII.4.3. GUARDAR DATOS Y GENERAR ARCHIVOS DE SALIDA

La exportación de datos es un proceso complementario de la importación. Las funciones para exportar datos generan archivos en un determinado formato. La estructura de la función es similar a las anteriores, con el prefijo «write» en lugar de «read», y con el nombre del archivo generado en lugar del archivo importado. Además se encuentran en los mismos paquetes.

Para formatos txt y csv tenemos «write.table()», «write.csv()» y «write.csv2()» (como se ha mencionado en el apartado anterior, el csv2 se distingue por utilizar coma para separar decimales y, punto y coma para separar campos). En todas estas funciones, el primer argumento entre paréntesis es el del objeto (matriz o *data frame*) que queremos enviar a nuestro nuevo archivo, y el segundo argumento, entrecomillado, es el nombre del archivo que generaremos, indicando la extensión. El archivo generado se graba en el directorio de trabajo que estemos utilizando.

Si, por ejemplo, quisiéramos escribir en archivo txt el conjunto de datos *iris*, tendríamos que introducir las siguientes instrucciones y luego ir al directorio de trabajo, donde encontraremos un nuevo archivo llamado **datairis.txt**.

```
> data(iris)
> write.table(iris,"datairis.txt")
```

Las funciones «write.csv()» y «write.csv2()» se utilizan de la misma manera. Por ejemplo, de nuevo importamos el archivo **Lagunas.csv,** pero ahora lo guardamos como **Lagunas2.csv** y le quitamos los códigos «NA» de las celdas en blanco con la instrucción «na=""». La instrucción «write.csv2()» es posiblemente la más adecuada para trabajar con archivos que estén configurados en español.

```
> datos<-read.csv2("Lagunas.csv",header=TRUE,encoding="latin1")
> write.csv2(datos,"Lagunas2.csv",na="")
```

	Laguna	Estación	Temperatura	Conductividad	pH	Nitrato	Nitrito	Amonio	Fosfato	Silicato
1	1	1	29,1	495	6,08	0	1,09	1,14	0,32	62,78
2	2	1	28,1	1537	7,85	0,52	0,28	0,94	0,384	231,673
3	2	2	28,6	1568			0,253	0,839	0,428	262,306
4	3	1	28,2	755	8,52	0,52	0,275	0,937	0,384	231,673
5	3	2	28,5	739	8,24	0	0,348	0,339	0,761	172,686
6	3	3	28,6	748	8,6	0,13	0,386	0,419	0,867	189,409
7	4	1	27,5	905	7,63	0	0,893	0,915	1,056	91,699
8	5	1	26,4	419	7,72		0,708	0,668	0,258	113,179
9	6	1	23,9	1034	7,64	0,42	0,301	1,229	0,263	189,038
10	7	1	32,6	217	7,42	0,3	0,843	2,076	0,618	117,996
11	8	1	31,3	371	7,32	0	0,697	0,635	0,489	99,225
12	9	1	32,3	1162	7,17	0,04	0,248	0,762	1,811	326,634

En el paquete "xlsReadWrite" existe también una función para generar archivos en Excel, «write.xls()», que se utiliza de un modo muy similar a las anteriores. Puede presentar problemas con los decimales cuando los archivos están configurados en español. Veamos un ejemplo con el conjunto de datos «iris»:

```
> library(xlsReadWrite)
> data(iris)
> write.xls(iris,"datairis.xls")
```

También es posible grabar archivos con el formato xlsx como se muestra en las siguientes instrucciones aunque, de nuevo, pueden existir problemas con el archivo grabado si la configuración está en español.

```
> library(xlsx)
> data(iris)
> write.xlsx(iris,"datairis.xlsx")
```

Existen más funciones de este tipo, especialmente en los paquetes "foreign" y "RODBC", para formatos menos habituales.

Por último, además de guardar datos, a la hora de programar en R resulta a menudo necesario crear un archivo que recoja todas las salidas de cada una de las pruebas que vamos haciendo. Para ello se utiliza la función «sink», en la cual debemos indicar dentro del paréntesis el nombre del archivo de salida y su extensión. Entonces se generará el archivo correspondiente en el directorio de trabajo, que recogerá todos los resultados que vayamos creando. Este archivo recogerá aquellos resultados que se muestren desde que la función se abre con el nombre del archivo de salida, hasta que indiquemos que la queremos detener, para lo cual basta con teclear el nombre de la función con los paréntesis vacíos «sink()». Veremos un ejemplo utilizando el conjunto de datos *iris*.

```
> data(iris)
> attach(iris)
> SetosaN<-shapiro.test(iris[(Species=="setosa"),"Petal.Length"])
> VirginicaN<-shapiro.test(iris[(Species=="virginica"),"Petal.Length"])
> Test<-t.test(iris[(Species=="setosa"),"Petal.Length"],
iris[(Species=="virginica"),"Petal.Length"])
> regr<-lm(Sepal.Length~Petal.Length)
> modelo<-summary(regr)
> Resultados<-list(SetosaN,VirginicaN,Test,modelo)
> Salida<-"Archivo de salida.txt"
> sink(Salida)
> print(Resultados)
> sink()
```

En el ejemplo anterior se realiza un test de Shapiro-Wilk para las especies *setosa* y *virginica*, a la variable «Petal.Length». Luego se hace un test de la *t* para determinar si existen diferencias en la variable «Petal.Length» entre ambas especies. Posteriormente, se realiza un análisis de regresión entre las variables «Sepal.Length» y «Petal.Length», considerando todas las especies. Por último, se guardan los resultados de todas las pruebas en **Archivo de salida.txt**., el cual se muestra en la siguiente ventana.

```
[[1]]

        Shapiro-Wilk normality test

data:  iris[(Species == "setosa"), "Petal.Length"]
W = 0.955, p-value = 0.05481


[[2]]

        Shapiro-Wilk normality test

data:  iris[(Species == "virginica"), "Petal.Length"]
W = 0.9622, p-value = 0.1098


[[3]]

        Welch Two Sample t-test

data:  iris[(Species == "setosa"), "Petal.Length"] and iris[(Species == "virginica"), "Petal.Length"]
t = -49.9862, df = 58.609, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.253749 -3.926251
sample estimates:
mean of x mean of y
    1.462     5.552


[[4]]

Call:
lm(formula = Sepal.Length ~ Petal.Length)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24675 -0.29657 -0.01515  0.27676  1.00269

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.30660    0.07839   54.94   <2e-16 ***
Petal.Length  0.40892    0.01889   21.65   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4071 on 148 degrees of freedom
Multiple R-squared:  0.76,      Adjusted R-squared: 0.7583
F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

AII.4.4. OBJETOS

La creación de objetos es la característica fundamental del lenguaje R. Los objetos se crean al ser asignados, y sus propiedades –pueden ser una función, el resultado de un test, etc.– se guardan en la memoria de la sesión en la que estemos trabajando. A partir de estos objetos también se pueden crear estructuras mayores, como las que veremos en los próximos apartados (funciones, vectores, matrices…).

Los nombres de objetos pueden contener letras mayúsculas o minúsculas, números y algunos signos de puntuación que no tengan un significado en R. Por ejemplo, no se podrían utilizar espacios en blanco, ni símbolos como: +, -, *, /, %, $, y otros.

La lista de objetos guardados en la memoria puede obtenerse con cualquiera de estas dos instrucciones:

```
> objects()
> ls()
```

Los objetos pueden eliminarse con la instrucción:

```
> rm(nombre del objeto)
```

o si deseamos eliminar todos los objetos almacenados en la memoria:

```
> rm(list=ls())
```

Un objeto en R posee una serie de atributos tales como clase, modo y estructura, los cuales es posible conocer mediante las siguientes instrucciones, respectivamente:

```
> class(nombre del objeto)
> mode(nombre del objeto)
> str(nombre del objeto)
```

La clase «class» puede ser «logical», «integer», «complex», «character», «data.frame», etc. Es atribuida por defecto, pero se puede asignar (en algunos casos) por medio de una instrucción de los siguientes tipos:

```
> as.logical(nombre del objeto)
> as.numeric(nombre del objeto)
> as.integer(nombre del objeto)
> as.character(nombre del objeto)
> as.matrix(nombre del objeto)
```

El modo «mode» se refiere a la forma que toma el objeto cuando es guardado en la memoria, y puede ser de tipo «logical», «integer», «complex», «character», en referencia a su clase o «list», «expression», «name», etc.

La estructura «str» indica diversos atributos pertenecientes al objeto, en formato de resumen compactado.

Veremos un ejemplo de creación de objetos simples y el manejo de todas las características anteriormente mencionadas. Para ello utilizaremos algunos vectores y matrices sencillos. Aunque todavía no hemos explicado cómo se construyen estos objetos, resultan adecuados para el ejemplo y fáciles de entender.

Comenzamos creando un objeto que será un vector de números impares al que llamamos «vector1» (la letra «c», antepuesta al paréntesis, indica que lo que viene a continuación es un vector) y si tecleamos el nombre de este nuevo objeto se nos devuelve su valor:

```
> vector1<-c(1,3,5,7,9)
> vector1
[1] 1 3 5 7 9
```

si queremos saber la clase de este objeto, se nos indica que es numérico:

```
> class(vector1)
[1] "numeric"
```

si ahora nos interesa saber el modo en que está guardado, se observa que en la memoria está almacenado en modo numérico, que coincide con su clase:

```
> mode(vector1)
[1] "numeric"
```

si ahora quisiéramos conocer su estructura, este objeto está compuesto por 5 números, cuyos valores también se indican:

```
> str(vector1)
 num [1:5] 1 3 5 7 9
```

Ahora crearemos otro objeto que será un vector del mismo tamaño, pero cuyos componentes serán letras (nótese cómo las letras se entrecomillan, ya que si no fuese así, R interpretaría que son objetos ya definidos):

```
> vector2<-c("a","b","c","d","e")
> vector2
[1] "a" "b" "c" "d" "e"
> class(vector2)
[1] "character"
> mode(vector2)
[1] "character"
> str(vector2)
 chr [1:5] "a" "b" "c" "d" "e"
```

A continuación crearemos un objeto, con 2 dimensiones (una matriz) uniendo ambos elementos:

```
> matriz1<-rbind(vector1,vector2)
> matriz1
        [,1] [,2] [,3] [,4] [,5]
vector1 "1"  "3"  "5"  "7"  "9"
vector2 "a"  "b"  "c"  "d"  "e"
> class(matriz1)
[1] "matrix"
> mode(matriz1)
[1] "character"
> str(matriz1)
 chr [1:2, 1:5] "1" "a" "3" "b" "5" "c" "7" "d" "9" "e"
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:2] "vector1" "vector2"
  ..$ : NULL
```

Como se observa en la ventana anterior, es una matriz y se encuentra en la memoria en modo carácter. Por lo tanto, los números, al fusionarse con letras, han pasado a considerarse caracteres. Esto tiene sentido, ya que R debe escoger entre el tipo «numeric», numérico del «vector1», o el tipo «character», carácter del «vector2», y el único que sería coherente dentro de un mismo objeto sería «character».

En la estructura se nos indica que nuestra matriz tiene dos filas y 5 columnas, cuyos elementos se listan ordenados. Además, se nos dice que la dimensión filas tiene dos nombres, formados por los caracteres «vector1» y «vector2», mientras que las columnas no tienen nombre.

Podemos cambiar la clase de alguno de los objetos creados. Por ejemplo, si queremos considerar a «vector1», que es de clase numérica, como categórico, podemos crear un nuevo objeto con clase «character»:

```
> vector3<-as.character(vector1)
> vector3
[1] "1" "3" "5" "7" "9"
```

Si intentamos cambiar «vector2» de clase, creando un objeto nuevo, vemos que el resultado no es muy satisfactorio, ya que no hay forma de adjudicar valores numéricos a categorías, y nos devuelve en su lugar el valor «NA», del que hablaremos en el próximo apartado, que indica un valor perdido. Por lo tanto, en este intento de transformación, se nos devolvería un vector sin datos.

```
> vector4<-as.numeric(vector2)
Mensajes de aviso perdidos
NAs introducidos por coerción
> vector4
[1] NA NA NA NA NA
```

Vamos ahora a comprobar cuántos objetos hemos ido almacenando en la memoria con la instrucción «objects()» y, por último, se borran todos los objetos de la memoria con «rm(list=ls())»:

```
> objects()
[1] "vector1" "vector2" "vector3" "vector4"
> rm(list=ls())
```

AII.4.5. OPERADORES

Se utilizan para operar con números y objetos. La tabla AII.1 muestra algunos tipos de operadores. Tecleando órdenes del tipo de las siguientes, se nos envía a una página de ayuda en HTML donde se muestran operadores y funciones genéricas de grupo:

```
>?Arith
>?Compare
>?Logic
```

Tabla AII. 1. Diferentes tipos de operadores que se usan en R.

Aritméticos	+	Suma
	-	Resta
	*	Multiplicación
	/	División
	^	Potencia
	%%	Módulo
	%/%	División en enteros
Lógicos	&	y
	&&	y
	|	o
	||	o
	!	No
	==	Igual
	<=	Menor o igual que
	>=	Mayor o igual que
	!=	Diferente de
	TRUE	Verdadero
	FALSE	Falso
Otros	NA	Dato ausente
	NaN	No es un número
	Inf	Infinito positivo
	-Inf	Infinito negativo
	~	Fórmula de un modelo
	$	Indexador de elementos
	:	Genera secuencia

AII.4.6. FUNCIONES

AII.4.6.1. Funciones *built in*

Se denomina así a aquellas funciones que ya se encuentran programadas en R, con lo cual sólo es necesario saber su nombre para poder utilizarlas. Existen

cientos de funciones. En la tabla AII.2 se muestran unos cuantos ejemplos entre las presentes en el paquete "base".

De la misma forma que con los operadores, es posible buscar información sobre funciones genéricas de grupo tecleando:

```
>?Math
>?Math2
>?Summary
>?Complex
```

Tabla AII.2. Diferentes tipos de funciones presentes en R *base*.

Tipo	Función	Uso
Aritméticas	log(x)	Logaritmo en base *e* de *x*
	log(x,n)	Logaritmo en base *n* de *x*
	log10(x)	Logaritmo en base 10 de *x*
	exp(x)	Antilogaritmo de *x* (e^x)
	sqrt(x)	Raíz cuadrada de *x*
	factorial(x)	*x!*
	choose(n,x)	Coeficiente binomial *n!/(x!)(n−x!)*
	ceiling(x)	Menor entero ≥ *x*
	floor(x)	Mayor entero ≤ *x*
	gamma(x)	*Γ(x)* para *x* real; *(x-1)!* Para *x* entero
	lgamma(x)	Logaritmo natural de *Γ(x)*
	trunc(x)	Entero más cercano entre *x* y 0
	round(x)	Redondeo a un entero
	signif(x)	Notación científica en 6 dígitos
	runif(n)	Genera *n* números aleatorios entre 0 y 1 a partir de una distribución uniforme
	abs(x)	Valor absoluto de *x*
Trigonométricas	sin(x)	Seno de *x* en radianes
	cos(x)	Coseno de *x* en radianes
	tan(x)	Tangente de *x* en radianes
Complejas	Arg	Ángulo en radianes del número complejo
	Conj	Conjugado del complejo
	Im	Parte imaginaria del número complejo
	Re	Parte real del número complejo
	Mod	Módulo
Conversión atributos de objetos	as.numeric(x) as.character(x) as.integer(x) …	Convierte la clase de *x* en la que se indica

Propiedades de objetos	is.numeric(x) is.character(x) is.na(x) …	Muestra una respuesta (TRUE o FALSE) acerca del tipo del objeto
Resúmenes	max(x)	Valor máximo
	min(x)	Valor mínimo
	range(x)	rango
	sum(x)	suma
	prod(x)	producto
Estadísticas, modelos, distribuciones	lm(formula)	Modelo lineal general
	glm(formula, family))	Modelo lineal generalizado
	nls(formula)	Regresión no lineal por mínimos cuadrados
	aov(formula)	Análisis de varianza
	rnorm(n,mean=0,sd=1)	Genera *n* números aleatorios extraídos de una distribución normal
	mean(x)	Media
	median(x)	Mediana
	var(x)	Varianza
	sd(x)	Desviación estándar
Gráficas	plot(x)	Gráfico de dispersión
	hist(x)	Histograma
	boxplot(x)	Diagrama de cajas
	pie (x)	Diagrama de tarta
Otras	na.omit(x)	Omitir valores perdidos
	na.action(x)	Tipo de acción que se efectúa al encontrar un valor perdido

Cada paquete aporta nuevas funciones, en algunos casos similares a las existentes en otros paquetes. También es posible que dos funciones con el mismo nombre en distintos paquetes hagan cosas diferentes. Se encuentra información sobre las funciones de un paquete en los manuales de cada paquete, en la página del proyecto R. Además, tenemos la denominada *R reference card* con algunas de las funciones más útiles en http://cran.r-project.org/doc/contrib/Short-refcard.pdf

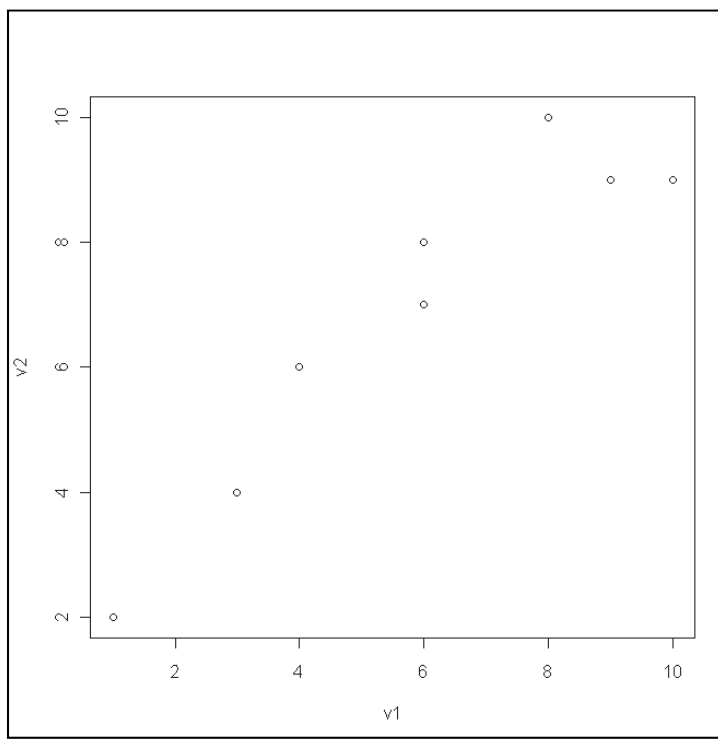
Todas las funciones tienen **argumentos** que se indican entre paréntesis después del nombre de la función. Estos argumentos pueden ser, por un lado, los objetos sobre los que se aplica dicha función y, por otro, una serie de características que es posible especificar para definir la forma en que actúa la función. Las funciones *built in*, por lo general, tienen unos pocos argumentos que es imprescindible indicar (el objeto sobre el que se aplica, y tal vez algunos parámetros) y el resto toman valores por defecto a menos que se especifique lo contrario. La mejor manera de saber qué argumentos deben utilizarse en una función es consultar la ayuda en HTML por medio de la instrucción:

>help(*nombre de la función*)

He aquí algunos ejemplos. En primer lugar, veremos uno con la función gráfica «plot». Creamos dos vectores que llamaremos «v1» y «v2», y vemos lo

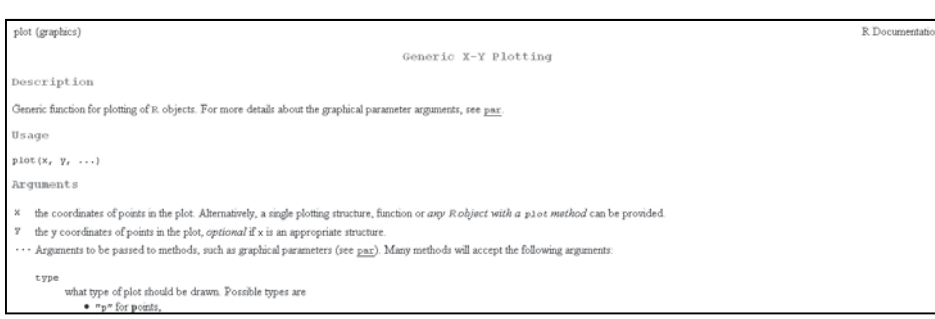
que ocurre si representamos gráficamente uno respecto del otro con la función «plot»:

```
> v1<-c(1,3,4,6,6,8,9,10)
> v2<-c(2,4,6,7,8,10,9,9)
> plot(v1,v2)
```



Si buscamos en la ayuda html las características de la función «plot», veremos qué argumentos es posible añadir para modificar su funcionamiento.

>help(plot)


plot (graphics) R Documentation

Generic X-Y Plotting

Description

Generic function for plotting of R objects. For more details about the graphical parameter arguments, see par.

Usage

plot(x, y, ...)

Arguments

x the coordinates of points in the plot. Alternatively, a single plotting structure, function or *any R object with a* plot *method* can be provided.

y the y coordinates of points in the plot, *optional* if x is an appropriate structure.

... Arguments to be passed to methods, such as graphical parameters (see par). Many methods will accept the following arguments:

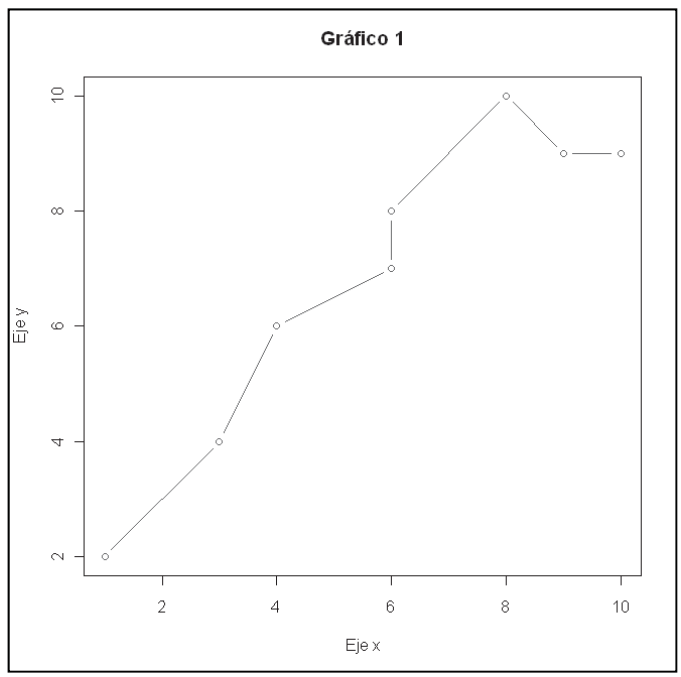
type
what type of plot should be drawn. Possible types are
- "p" for points,

Vemos que podemos añadir a la función argumentos para modificar el tipo de símbolos empleados en la representación, su color, añadir un título al gráfico,

añadir leyendas a los ejes, etc. Vamos entonces a acometer algunas de estas modificaciones:

```
> plot(v1,v2,type="b",col="red",main="Gráfico 1",xlab="Eje x",ylab="Eje y")
>
```

Ahora estaríamos diciendo lo siguiente: «type» es el argumento que indica el tipo de símbolo a usar, la «b» representa línea y puntos simultáneamente; «col» indica el color, en este caso rojo; «main», «xlab» y «ylab» son el título del gráfico y de los ejes de coordenadas. Este es el nuevo gráfico que se obtiene:



Es importante destacar que los valores que se le dan a los argumentos, así como el texto de los títulos, se entrecomillan, de modo que R no los confunde con un objeto.

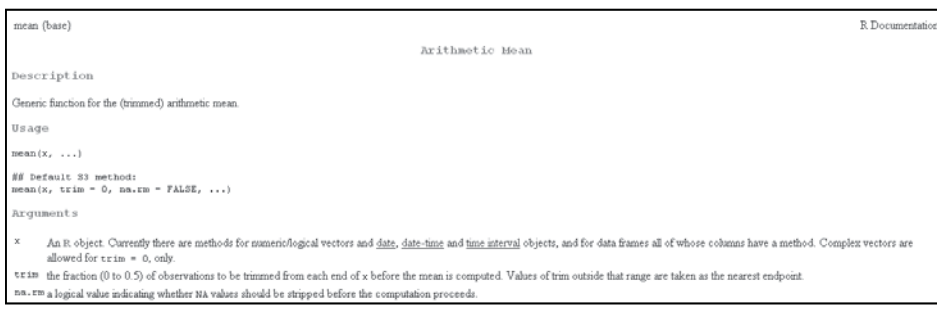
Veamos otro ejemplo con una función de tipo matemático-estadístico, como «mean». Empezaremos cargando un conjunto de datos de los que ya disponemos en R, como por ejemplo «iris», que se encuentra en el paquete "datasets", presente en la instalación por defecto. Lo que haremos será calcular la media de la variable «Sepal.Length»:

```
> data(iris)
> attach(iris)
> mean(Sepal.Length)
[1] 5.843333
```

Como explicamos en apartados anteriores, la instrucción «data» sirve para cargar un conjunto de datos que vienen con los paquetes. Además, con la instrucción «attach», adjuntamos este conjunto de datos en la memoria de forma

que simplemente con escribir el nombre de cada variable, R la busca en el conjunto de datos adjunto (si no, habría que indicarle que esa variable se encuentra en ese conjunto de datos). La media de todas las observaciones de longitud de sépalos es 5,84. Vamos ahora a consultar en la ayuda HTML qué otros argumentos es posible utilizar en la función «mean»:

>help(mean)


mean (base) R Documentation

Arithmetic Mean

Description

Generic function for the (trimmed) arithmetic mean.

Usage

```
mean(x, ...)

## Default S3 method:
mean(x, trim = 0, na.rm = FALSE, ...)
```

Arguments

x	An R object. Currently there are methods for numeric/logical vectors and date, date-time and time interval objects, and for data frames all of whose columns have a method. Complex vectors are allowed for trim = 0, only.
trim	the fraction (0 to 0.5) of observations to be trimmed from each end of x before the mean is computed. Values of trim outside that range are taken as the nearest endpoint.
na.rm	a logical value indicating whether NA values should be stripped before the mean proceeds.

Vemos que dos argumentos posibles para «mean» son: 1) especificar si se desea la media sobre el total de los datos o la media acotada, y en este último caso se debería especificar en qué fracción se acota (el acotamiento es por los dos extremos del rango de datos). Por defecto no se utiliza ninguna acotación «trim=0»; 2) especificar si se deben eliminar los valores perdidos, que por defecto no se eliminan «na.rm=FALSE». Los puntos suspensivos, además, indican que es posible añadir más argumentos que sean aplicables a esta función.

Así pues, nosotros vamos a realizar la media con un truncamiento del 0,1, y no eliminamos los valores perdidos (dejamos la opción por defecto) porque sabemos que en este conjunto de datos no existen valores perdidos. El valor de la media cambia ligeramente, haciéndose menor.

```
> mean(Sepal.Length, trim=0.1)
[1] 5.808333
>
```

AII.4.6.2. Funciones definidas por el usuario

Además de todas las funciones existentes en R, es posible programar otras nuevas para realizar cualquier operación que nos interese. Simplemente tenemos que ceñirnos al siguiente formato:

>*nombre*<-function(*lista de argumentos*){*expresión*}

Estas son algunas de las características acerca de cómo trabaja R con las funciones definidas por usuario:

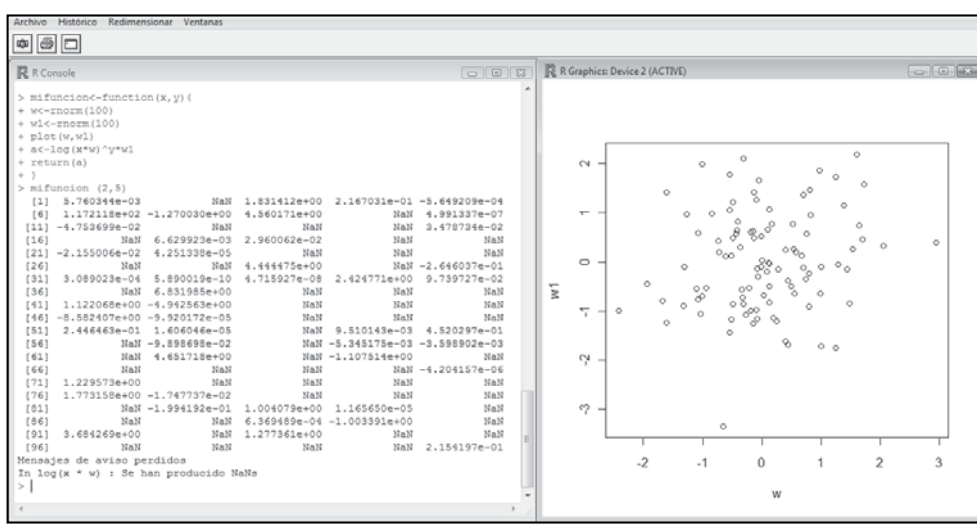
- Se pueden asignar valores a los parámetros en la definición de la función. Si no, los valores de los parámetros se especificarán al aplicarla.
- La función devuelve aquello que indiquemos con la instrucción «return», o si no, el último elemento evaluado.

- Si incluimos puntos suspensivos en la lista de argumentos, eso quiere decir que se pueden añadir nuevos argumentos a la hora de aplicarla. Esto es así también en las funciones *built in*, como ya vimos.
- El valor que nos devuelven las funciones se encuentra en modo de lista (veremos lo que es una lista en próximos apartados).

Aquí mostramos algunos ejemplos sencillos de creación de funciones y su aplicación:

```
> mifuncion<-function(x,y){
+             w<-rnorm(100)
+             w1<-rnorm(100)
+             plot(w,w1)
+             a<-log(x*w)^y*w1
+             return(a)
+ }
```

Primero le hemos puesto un nombre «mifuncion», y hemos escrito que esa será una función aplicable a dos valores sin especificar, «x» e «y». La expresión consiste en crear dos series de 100 números aleatorios extraídos de una distribución normal de media 0 y desviación típica 1, llamadas «w» y «w1». Posteriormente se representa una gráfica de dispersión de «w» frente a «w1» (que se nos mostrará ya que no la hemos almacenado en ningún objeto) y se realiza una operación con los valores de «x», «y», «w» y «w1», a cuyo resultado llamaremos «a», que es el valor que se nos devuelve (instrucción «return»). A la hora de escribir funciones es muy útil, aunque no necesario, escribir las instrucciones en diferentes niveles acorde con la estructura de la expresión que estamos creando. En el ejemplo de «mifuncion», esto se refleja al situar a la misma altura «w», «w1», «plot» y «a», ya que la expresión de cada uno es un módulo dentro de la función. Las llaves vemos que se cierran al nivel del nombre de la función. Ahora, si aplicamos esta función sobre dos números cualesquiera «x» e «y» con la instrucción «mifuncion (2,5)», obtendremos lo siguiente:



Como vemos, se ha representado el gráfico de dispersión y se muestra el valor de «a», que es una serie de 100 números, entre los cuales, en nuestro caso, hay valores –logaritmos de números negativos– que no es posible calcular, y se muestran como «NaN». Otro ejemplo de función definida por el usuario:

```
> mifuncion2<-function(x){
+              v1<--x
+              t1<-rnorm(1000)
+              for(i in t1){if(i>2)s1<-v1*t1^2
+                           else(s1<-v1*t1^3)
+                           }
+              return(s1)
+ }
>
```

En esta función, a la que hemos llamado «mifuncion2», creamos un objeto «v1» que es simplemente el valor de «x» cambiado de signo, otro objeto «t1» que son 1000 números aleatorios extraídos de una distribución normal de media 0 y desviación típica 1. Después hacemos un bucle, por medio de la función «for» sobre los valores de «t1» (explicaremos en qué consiste y como se hacen bucles en el apartado siguiente) en el cual incluimos una expresión lógica: si el valor del elemento de «t1» es mayor que 2, el valor del objeto «s1» se obtiene de la expresión «v1*t1^2», para todos los demás (es decir, para los valores menores o iguales a 2) el valor de «s1» se obtiene de la expresión «v1*t1^3». El valor que se nos devuelve de esta función es «s1».

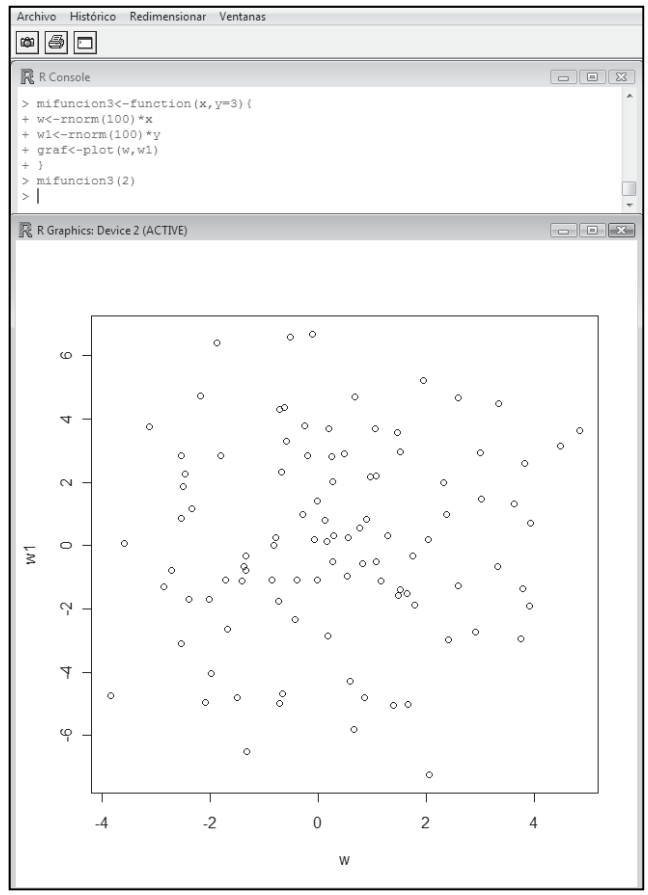
En la siguiente ventana se observa qué es lo que se obtiene al aplicar esta función para x = 4, es decir «mifuncion2(4)». Como vemos se trata de 1000 valores (no se muestran todos en esta captura de pantalla) obtenidos conforme a lo especificado en la función.

```
> mifuncion2(4)
   [1] -2.385656e-01 -1.707230e+00 -5.273684e-02  1.088036e-01  4.492563e-01
   [6] -5.832089e+00  7.316737e+00 -2.071078e+00 -2.232748e-03  8.746299e-04
  [11] -3.920089e-04 -5.659597e-01  3.107056e-03  2.423545e-03 -6.050802e+00
  [16]  2.401833e+01 -1.853385e-01  4.970294e+00  1.726044e+01 -1.434039e+01
  [21]  5.718472e+00 -1.551005e-02 -2.222725e-04 -1.469450e+00 -7.180048e+00
  [26] -7.633437e+00 -1.547317e+01  2.824386e+00 -8.430173e+00 -1.615950e-01
  [31]  8.934756e+01  1.308790e-01 -4.842051e+00  4.021506e+00 -7.685303e+00
```

Basándonos en ejemplos sencillos como los anteriores, vamos a ver más posibilidades en torno a la creación de funciones:

```
> mifuncion3<-function(x,y=3){
+              w<-rnorm(100)*x
+              w1<-rnorm(100)*y
+              graf<-plot(w, w1)
+ }
```

En esta nueva función hemos definido al parámetro «y» con un valor fijo. Además, no hemos incluido la instrucción «return». Esto es lo que se obtiene al aplicar la instrucción «mifuncion3(2)»:

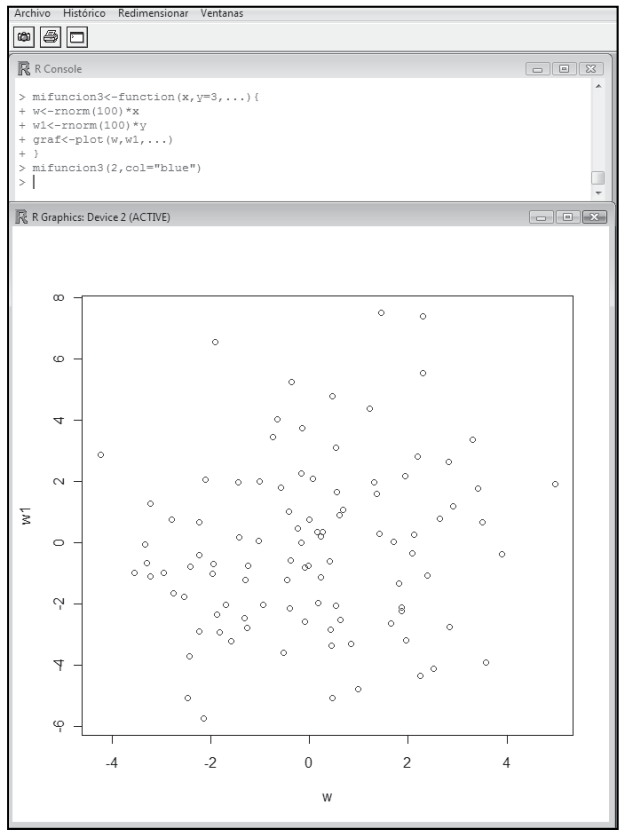


Vemos cómo no es necesario especificar el valor de «y», ya que está establecido como 3, y cómo, ya que no se especifica nada por medio de la instrucción «return», se nos devuelve simplemente el último objeto evaluado, en este caso «graf».

Ahora haremos una pequeña modificación en la función anterior, de modo que exista la posibilidad de incluir argumentos sin definir previamente:

```
> mifuncion3<-function(x,y=3,...){
+             w<-rnorm(100)*x
+             w1<-rnorm(100)*y
+             graf<-plot(w, w1,...)
+ }
```

La función es la misma de antes, pero se colocan puntos suspensivos en la lista de argumentos inicial, de modo que sea posible añadirlos a la hora de aplicar la función. Además, como se dejan puntos suspensivos también para «plot», dichos argumentos podrían pertenecer a la función «plot». Se introduce la instrucción «mifuncion3(2,col="blue")»:



Lo que hemos hecho es incluir un argumento «col» que es específico de las funciones para graficar, y que nos indica de qué color deben representarse los símbolos en nuestra gráfica.

AII.4.7. BUCLES, CICLOS O *LOOPS*

Un bucle, ciclo o loop en programación, consiste en aplicar un código sobre un número de elementos de forma que se repite en cada elemento la operación definida en ese código. Los bucles en R pueden hacerse con las funciones «for» (la más común), «while» y «repeat». En realidad, instrucciones de la familia «apply» («apply», «sapply», «tapply», «mapply», «lapply») también ejecutan bucles, pero los veremos más adelante. Los bucles se efectúan siempre sobre vectores u objetos que puedan ser tratados como tales.

Los bucles con «for» tienen la siguiente estructura. Entre paréntesis se nos indica sobre qué elementos de una determinada expresión (dicha expresión puede ser cualquier vector de longitud distinta de 0) se va a aplicar el bucle y después, entre llaves, se indica la operación a realizar. He aquí un ejemplo:

```
> for(i in 1:20){
+      a<-log(i)
+      b<-a*i
+      print(b)
+ }
```

Primero, creamos un secuencia de enteros entre 1 y 20 «1:20» sobre la que se aplicará la operación del bucle, en cada elemento iésimo de esa secuencia «i in 1:20». Dentro de la expresión que se aplicará sobre cada elemento de la secuencia tenemos la creación de dos objetos «a» y «b», de los cuales imprimimos «b» con «print». El resultado es el siguiente:

```
[1] 0
[1] 1.386294
[1] 3.295837
[1] 5.545177
[1] 8.04719
[1] 10.75056
[1] 13.62137
[1] 16.63553
[1] 19.77502
[1] 23.02585
[1] 26.37685
[1] 29.81888
[1] 33.34434
[1] 36.9468
[1] 40.62075
[1] 44.36142
[1] 48.16463
[1] 52.02669
[1] 55.94434
[1] 59.91465
>
```

También es posible hacer bucles con la función «while». En este caso la estructura que se utiliza es distinta. La expresión entre paréntesis establece una condición que debe ser evaluada, y si es cierta, entonces se aplica la expresión que viene a continuación (habitualmente entre llaves) después se vuelve a evaluar la expresión entre paréntesis y si todavía es cierta se vuelve a aplicar la expresión entre llaves. Así sucesivamente hasta que la expresión entre paréntesis sea falsa. Ejemplo, dentro de una función:

```
> func1<-function(x) {
+ a <- 1
+ b <- x
+ while(b<10) {
+ a <- a*b
+ b <- b+5 }
+ return(a) }
> sapply(3:7,func1)
[1] 24 36  5  6  7
```

En primer lugar hemos definido la función «func1», dentro de la cual está el bucle con «while». La expresión evaluada con el bucle es que mientras «b» sea menor que 10, «b» pasa a ser «b + 5», y «a» se recalcula utilizando el anterior valor modificado de «a». Una vez que «b» pasa a ser igual o mayor que 10, ya no se realizan más operaciones y la función devuelve el último valor calculado de «a». Este mecanismo se observa claramente si aplicamos esta función a la serie de números de 3 a 7, como se puede observar en el último comando.

Otro tipo de bucle menos común es el que se hace con «repeat». En este caso la estructura consiste en que se repita una instrucción hasta que se cumpla una determinada condición, que debe señalarse con «break», de lo contrario el bucle

sería infinito. En el siguiente ejemplo se observa como readaptando para «repeat» el ejemplo anterior de «func1», se produce una función con idéntico resultado:

```
> func2<-function(x){
+ a <- 1
+ b <- x
+ repeat{
+ if (b>=10)break
+ a <-a*b
+ b <- b+5}
+ return(a)}
> sapply(3:7,func2)
[1] 24 36  5  6  7
```

AII.4.8. VECTORES

Una característica importante del lenguaje de R es que es vectorial en sus operaciones. Esto quiere decir que las operaciones matemáticas se realizan directamente con los vectores, sin que sea necesario programar operaciones entre sus elementos.

AII.4.8.1. Creación de vectores

Los vectores son variables con uno o más elementos del mismo tipo (números, caracteres, valores lógicos, etc.). Hay varias formas de generar vectores. Una forma es escribiendo directamente los elementos, usando la función de concatenación «c»:

```
>vector1<-c(2,4,6,8,3,2,5,6,7)
>vector2<-c("a","d","f","g","t","e","c","v","g","y","u")
>vector3<-c(TRUE,FALSE,FALSE,TRUE,FALSE,TRUE,FALSE,FALSE)
```

Otra forma es generando secuencias de números enteros con el operador «**:**»:

```
>vector1<-1:12
```

Existen funciones específicas para generar secuencias. Los argumentos de «seq» son el inicio de la secuencia, el fin, y el paso. Así podemos crear un vector:

```
>seq(0,5,0.5)
```

```
> seq(0,5,0.5)
 [1] 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
```

Otra función es «sequence», en la que cada elemento es una secuencia:

>sequence(c(3,2,5))

```
> sequence(c(3,2,5))
 [1] 1 2 3 1 2 1 2 3 4 5
```

También es posible generar vectores con repeticiones, para esto existe la función «rep», en la cual el primer argumento es el número o números que se repetirán, el segundo «each» cuántas veces se repetirán, y el tercero «times» indica cuántas veces se repite la estructura total (este argumento no es imprescindible). Por ejemplo:

>rep(c(2,5,3),each=3,times=2)

```
> rep(c(2,5,3),each=3,times=2)
 [1] 2 2 2 5 5 5 3 3 3 2 2 2 5 5 5 3 3 3
```

Si usamos menos argumentos, no es necesario escribir el nombre, y R los interpreta bien:

>v2<-rep(5,8)
>v2

```
> v2<-rep(5,8)
> v2
[1] 5 5 5 5 5 5 5 5
```

Existe un vector *built in* llamado «letters», si queremos utilizar minúsculas, o «LETTERS», si queremos mayúsculas. Contiene todas las letras en orden alfabético:

>letters

```
> letters
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m" "n" "o" "p" "q" "r" "s" "t" "u" "v" "w" "x" "y" "z"
```

Los vectores cuyos componentes son caracteres, automáticamente se consideran como variables categóricas, y deben señalarse con comillas. También nos puede interesar generar vectores de variables categóricas, para ser utilizados como factores. Para esto existe una función denominada «gl», de *generate levels* (referido a niveles de un factor), que tiene una estructura de argumentos similar a «seq»; entre paréntesis deben indicarse: número de niveles, número de repeticiones y longitud total. Como se observa, siempre que tengamos un objeto como factor, se nos indica al final el número total de niveles del factor «Levels: 1 2 3 4».

>mifactor<-gl(4,5,30)
>mifactor

```
> mifactor<-gl(4,5,30)
> mifactor
 [1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 1 1 1 1 1 2 2 2 2 2
Levels: 1 2 3 4
```

Es posible que nos interese nombrar de alguna forma a los elementos de la variable que tenemos dentro de un vector. Para esto se usa la función «names», que también veremos para matrices, listas y *data frames*. Veamos algunos ejemplos. En este primer caso, como nombres utilizaremos simplemente números enteros:

```
>v1<-c(21,25,36,24,52,42,63,58)
>names(v1)<-0:7
>v1
```

```
> v1<-c(21,25,36,24,52,42,63,58)
> names(v1)<-0:7
> v1
 0  1  2  3  4  5  6  7
21 25 36 24 52 42 63 58
```

pero podemos utilizar también categorías:

```
>names(v1)<-c("m1","m2","m3","m4","m5","m6","m7","m8")
>v1
```

```
> names(v1)<-c("m1","m2","m3","m4","m5","m6","m7","m8")
> v1
m1 m2 m3 m4 m5 m6 m7 m8
21 25 36 24 52 42 63 58
```

AII.4.8.2. Trabajando con vectores

Hay numerosas funciones que actúan sobre todos los elementos de un vector y nos devuelven un valor, algunas de las cuales se vieron en la tabla AII.2. A modo de ejemplo:

```
>v1<-c(10,25,77,8,14,36)
>mean(v1)
>length(v1)
>max(v1);min(v1)
```

```
> v1<-c(10,25,77,8,14,36)
> mean(v1)
[1] 28.33333
> length(v1)
[1] 6
> max(v1);min(v1)
[1] 77
[1] 8
```

Otro tipo de funciones nos permite extraer elementos al azar, ordenarlos, etc. Veamos algunos ejemplos:

```
>v2<-c(21,25,36,24,52,42,63,58)
>v2
>sample(v2)
```

```
>rank(v2)
>order(v2)
>sort(v2)
```

```
> v2<-c(21,25,36,24,52,42,63,58)
> v2
[1] 21 25 36 24 52 42 63 58
> sample(v2)
[1] 58 25 36 21 24 52 42 63
> rank(v2)
[1] 1 3 4 2 6 5 8 7
> order(v2)
[1] 1 4 2 3 6 5 8 7
> sort(v2)
[1] 21 24 25 36 42 52 58 63
```

Como vemos, la función «sample» devuelve una secuencia al azar de los elementos del vector. La función «rank» nos asigna rangos a cada elemento de menor a mayor. La función «order» ordena los elementos (mostrando su nombre) de menor a mayor, la función «sort» ordena los elementos (mostrando su valor) de menor a mayor.

Una función importante, «sapply» (relacionada con otras que veremos más adelante: «apply», «lapply», «tapply») permite aplicar una función a cada uno de los componentes de un vector. Por ejemplo, si creamos un vector que es una secuencia de 1 a 5 y queremos calcular el factorial de cada número:

```
> sapply(1:5,factorial)
```

```
> sapply(1:5,factorial)
[1]   1   2   6  24 120
```

Una opción interesante de los vectores es evaluar su contenido mezclando expresiones lógicas y aritméticas. Por ejemplo, si queremos saber qué elementos de «v2» son mayores que 40, o si alguno o todos los elementos de «v2» son mayores que 40:

```
>v2>40
> any(v2>40)
> all(v2>40)
```

```
> v2>40
[1] FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE
> any(v2>40)
[1] TRUE
> all(v2>40)
[1] FALSE
```

Para localizar elementos de un vector se usan corchetes. Si queremos que se nos devuelva el valor de uno o varios elementos dentro de ese vector, simplemente escribimos la posición del elemento que queremos:

```
> v2[5]
> v2[c(2,5)]
```

```
> v2[5]
[1] 52
> v2[c(2,5)]
[1] 25 52
```

También es interesante localizar dentro de vectores, aquellos elementos que tienen un determinado valor. Esto se hace de forma sencilla utilizando la función «which». Por ejemplo, ver en qué posición de «v2» está el valor 36, o también se puede pedir un rango de valores que cumplan una condición:

```
>which(v2==36)
>which(v2<40)
```

```
> which(v2==36)
[1] 3
> which(v2<40)
[1] 1 2 3 4
```

Otra posibilidad importante para extraer componentes y modificar vectores es eliminar elementos, anteponiendo un símbolo «-» al número del elemento:

```
>v2<-c(21,25,36,24,52,42,63,58)
>v2
>v2[-3]
```

```
> v2<-c(21,25,36,24,52,42,63,58)
> v2
[1] 21 25 36 24 52 42 63 58
> v2[-3]
[1] 21 25 24 52 42 63 58
```

AII.4.9. MATRICES Y *ARRAYS*

Las matrices se interpretan en R como vectores con dos dimensiones (filas y columnas). Son objetos de clase *matrix*, y como atributos tienen sus dimensiones. Un *Array* es equivalente a una matriz, pero con cualquier número de dimensiones.

AII.4.9.1. Creación de Matrices y *Arrays*

La forma más elemental de crear una matriz, al igual que un vector, consiste en teclear sus elementos directamente. Pero aquí hay que prestar atención al orden en que se introducen, ya que por defecto se interpreta que los valores se introducen en sentido de columnas. Además hay que especificar al menos el valor de una dimensión (número de filas o número de columnas). Por ejemplo, queremos una matriz de 4 filas y 3 columnas, y en este caso indicamos el número de filas:

```
>A<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8), nrow=4)
>A
```

```
> A<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8), nrow=4)
> A
     [,1] [,2] [,3]
[1,]    2    0    7
[2,]    5    6    4
[3,]    6    5    9
[4,]    1    8    8
```

Observamos que a filas y columnas se les adjudica un número (ya que no hay ningún nombre específico). Las filas se consideran la primera dimensión, por eso su número se sitúa a la izquierda de la coma. Las columnas la segunda, por eso aparece a la derecha. Veremos que esto es muy importante. Si indicásemos el número de columnas, el resultado sería el mismo:

```
>A<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8), ncol=3)
>A
```

```
> A<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8), ncol=3)
> A
     [,1] [,2] [,3]
[1,]    2    0    7
[2,]    5    6    4
[3,]    6    5    9
[4,]    1    8    8
```

Es posible introducir los datos por filas, si lo indicamos con el argumento «byrow», y la matriz cambia por completo:

```
>A<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8), byrow=TRUE, ncol=3)
>A
```

```
> A<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8), byrow=TRUE, ncol=3)
> A
     [,1] [,2] [,3]
[1,]    2    5    6
[2,]    1    0    6
[3,]    5    8    7
[4,]    4    9    8
```

También puede resultar útil, en ciertos casos, convertir un vector en una matriz. Para ello se utiliza la función «dim», y hay que indicar las dimensiones que deseamos:

```
>v1<-c(1,0,1,0,0,1,1,1,1,1,0,0)
>dim(v1)<-c(3,4)
>v1
```

```
> v1<-c(1,0,1,0,0,1,1,1,1,1,0,0)
> dim(v1)<-c(3,4)
> v1
     [,1] [,2] [,3] [,4]
[1,]    1    0    1    1
[2,]    0    0    1    0
[3,]    1    1    1    0
```

En muchos casos necesitaremos nombrar las filas y/o columnas de una matriz. Lo más elemental sería crear un vector de longitud igual al número de nombres que necesitamos. Existen opciones más avanzadas, que veremos en el apartado referente a texto. Unos ejemplos con los métodos más sencillos serían con las funciones «rownames» y «colnames», retomando la matriz «A» del ejemplo anterior, en la que queremos indicar que los números corresponden a los valores obtenidos en una serie de análisis de pequeñas moléculas orgánicas en 4 muestras:

```
>A<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8), byrow=TRUE, ncol=3)
>A
>Pruebas<-c("Glu","Sac","Asp")
>colnames(A)<-Pruebas
>Muestras<-c("m1","m2","m3","m4")
>rownames(A)<-Muestras
>A
```

```
> A<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8), byrow=TRUE, ncol=3)
> A
     [,1] [,2] [,3]
[1,]    2    5    6
[2,]    1    0    6
[3,]    5    8    7
[4,]    4    9    8
> Pruebas<-c("Glu","Sac","Asp")
> colnames(A)<-Pruebas
> Muestras<-c("m1","m2","m3","m4")
> rownames(A)<-Muestras
> A
   Glu Sac Asp
m1   2   5   6
m2   1   0   6
m3   5   8   7
m4   4   9   8
```

Otra posibilidad es utilizar la función «dimnames», en la que se nombrarían filas y columnas en la misma expresión. Los nombres de las dimensiones deberán tener formato de lista (veremos en el próximo apartado qué es una lista) en la cual primero se indican filas y después columnas. En este ejemplo, de nuevo con la matriz anterior, no necesitamos nombrar las filas ya que sabemos que cada una equivale a una muestra, por lo que los números que tienen por defecto son adecuados. Ya que no queremos nombrar las filas, deberemos indicar «NULL» en el lugar en el que debería situarse el nombre de las filas:

```
>A<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8),byrow=TRUE,ncol=3)
>A
>Pruebas<-c("Glu","Sac","Asp")
>dimnames(A)<-list(NULL,Pruebas)
>A
```

```
> A<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8),byrow=TRUE,ncol=3)
> A
     [,1] [,2] [,3]
[1,]    2    5    6
[2,]    1    0    6
[3,]    5    8    7
[4,]    4    9    8
> Pruebas<-c("Glu","Sac","Asp")
> dimnames(A)<-list(NULL,Pruebas)
> A
     Glu Sac Asp
[1,]   2   5   6
[2,]   1   0   6
[3,]   5   8   7
[4,]   4   9   8
```

Otra opción interesante para configurar nuestras matrices es que se pueden añadir filas o columnas a una matriz ya creada, por medio de las funciones «rbind» y «cbind». En estas funciones, simplemente es necesario indicar cuál es la matriz a la que deseamos añadirle un componente y cuál es ese componente. Por ejemplo, añadiremos a la matriz «A», el resultado de una prueba más:

```
>A
>Act<-c(0,3,5,9)
>A<-cbind(A,Act)
>A
```

```
> A
     Glu Sac Asp
[1,]   2   5   6
[2,]   1   0   6
[3,]   5   8   7
[4,]   4   9   8
> Act<-c(0,3,5,9)
> A<-cbind(A,Act)
> A
     Glu Sac Asp Act
[1,]   2   5   6   0
[2,]   1   0   6   3
[3,]   5   8   7   5
[4,]   4   9   8   9
```

Y si queremos añadir los resultados del análisis de una muestra, como una nueva fila sin nombre:

```
>A<-rbind(A,c(0,2,5,10))
>A
```

```
> A<-rbind(A,c(0,2,5,10))
> A
     Glu Sac Asp Act
[1,]   2   5   6   0
[2,]   1   0   6   3
[3,]   5   8   7   5
[4,]   4   9   8   9
[5,]   0   2   5  10
```

Todos los ejemplos y funciones anteriores serían válidos para *Arrays*, que en realidad son también matrices con cualquier número de dimensiones. Por ejem-

plo, vamos a crear un *Array* sencillo, de tres dimensiones. Primero creamos un vector, que será una secuencia de enteros, desde 1 hasta 50:

```
>B<-c(1:50)
```

y ahora le damos tres dimensiones, en este caso será 5x5x2 (ya que B tiene 50 elementos):

```
>dim(B)<-c(5,5,2)
>B
```

```
> B<-c(1:50)
> dim(B)<-c(5,5,2)
> B
, , 1

     [,1] [,2] [,3] [,4] [,5]
[1,]    1    6   11   16   21
[2,]    2    7   12   17   22
[3,]    3    8   13   18   23
[4,]    4    9   14   19   24
[5,]    5   10   15   20   25

, , 2

     [,1] [,2] [,3] [,4] [,5]
[1,]   26   31   36   41   46
[2,]   27   32   37   42   47
[3,]   28   33   38   43   48
[4,]   29   34   39   44   49
[5,]   30   35   40   45   50
```

La tercera dimensión se muestra como varias matrices bidimensionales, y se indica su número o su nombre (si lo tuviera) a la derecha de la segunda coma. También es posible crear *Arrays* con la función «array»:

```
>array(1:50,dim=c(5,5,2))
```

```
> array(1:50,dim=c(5,5,2))
, , 1

     [,1] [,2] [,3] [,4] [,5]
[1,]    1    6   11   16   21
[2,]    2    7   12   17   22
[3,]    3    8   13   18   23
[4,]    4    9   14   19   24
[5,]    5   10   15   20   25

, , 2

     [,1] [,2] [,3] [,4] [,5]
[1,]   26   31   36   41   46
[2,]   27   32   37   42   47
[3,]   28   33   38   43   48
[4,]   29   34   39   44   49
[5,]   30   35   40   45   50
```

AII.4.9.2. Trabajando con Matrices y *Arrays*

Antes de empezar a hablar de operaciones y funciones que es posible aplicar a matrices, es necesario aprender a seleccionar componentes o elementos de una matriz. Esto es así porque, particularmente, muchas de las funciones que veremos operan sólo sobre una fila o una columna.

Al igual que con vectores, para señalar un componente de una matriz se utilizan corchetes «[]». Pero en este caso, como tenemos dos dimensiones dentro de una matriz, dentro del corchete hay que situar una coma, a la izquierda de la cual se indican las filas, y a la derecha las columnas «[,]». Cada dimensión se nombra con el número que le corresponde. Se puede utilizar también su nombre, en lugar del número, si lo tiene. Veamos ejemplos. Seleccionamos una fila, con los elementos de todas las columnas. Para ello debemos indicar el número de la fila a la izquierda de la coma y no indicar ningún número para las columnas:

```
>M<-matrix(c(1,1,1,0,0,1,0,1,1,0,1,1),nrow=3)
>M
>M[2,]
```

```
> M<-matrix(c(1,1,1,0,0,1,0,1,1,0,1,1),nrow=3)
> M
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    1    0    1    1
[3,]    1    1    1    1
> M[2,]
[1] 1 0 1 1
```

Para la misma matriz, si queremos seleccionar una columna:

```
>M[,4]
```

```
> M[,4]
[1] 0 1 1
```

Si queremos seleccionar tan sólo un elemento de la matriz, habría que indicar la fila y la columna:

```
>M[2,4]
```

```
> M[2,4]
[1] 1
```

Si queremos seleccionar dos filas:

```
>M[c(2,3),]
```

```
> M[c(2,3),]
     [,1] [,2] [,3] [,4]
[1,]    1    0    1    1
[2,]    1    1    1    1
```

Si queremos dos columnas, se hace de un modo similar:

> M[,c(1,4)]

```
> M[,c(1,4)]
     [,1] [,2]
[1,]    1    0
[2,]    1    1
[3,]    1    1
```

Es conveniente fijarse en que los números que tienen las filas y columnas seleccionadas no son los antiguos. También podemos seleccionar un elemento en una dimensión y varios en la otra. Por ejemplo, una fila y dos columnas:

> M[2,c(2,3)]

```
> M[2,c(2,3)]
[1] 0 1
```

Si en lugar de números, las dimensiones de nuestra matriz tuviesen texto, el proceso de selección sería similar. No obstante, es interesante reseñar que la selección utilizando números sigue siendo posible aún cuando las dimensiones tengan nombres, ya que en la memoria se les sigue asignando un número a las dimensiones:

```
>colnames(M)<-c("PruebaA","PruebaB","PruebaC","PruebaD")
>rownames(M)<-c("m1","m2","m3")
>M
>M["m2",]
>M[,"PruebaB"]
>M["m2",c("PruebaB","PruebaC")]
```

```
> colnames(M)<-c("PruebaA","PruebaB","PruebaC","PruebaD")
> rownames(M)<-c("m1","m2","m3")
> M
   PruebaA PruebaB PruebaC PruebaD
m1       1       0       0       0
m2       1       0       1       1
m3       1       1       1       1
> M["m2",]
PruebaA PruebaB PruebaC PruebaD
      1       0       1       1
> M[,"PruebaB"]
m1 m2 m3
 0  0  1
```

Al igual que hemos visto con vectores, en matrices es posible eliminar elementos si anteponemos un símbolo negativo al número del elemento:

```
>X<-matrix(c(1,1,0,1,0,1,0,1,1),nrow=3)
>X
>X[-2,-1]
```

```
> X<-matrix(c(1,1,0,1,0,1,0,1,1),nrow=3)
> X
     [,1] [,2] [,3]
[1,]    1    1    0
[2,]    1    0    1
[3,]    0    1    1
> X[-2,-1]
     [,1] [,2]
[1,]    1    0
[2,]    1    1
```

Existen numerosas funciones para hacer cálculos específicos de matrices. Por ejemplo «t(x)» calcula la matriz transpuesta de «x», «det(x)» calcula el determinante de «x», «diag (x)» extrae la diagonal de «x», «crossprod(x,y)» calcula el producto cruzado de «x» e «y». También tenemos operadores exclusivos para matrices: «%*%» es el producto matricial, «%x%» el producto de Kronecker, entre otros. Veamos algunos ejemplos de todo esto, por ejemplo, el producto matricial:

```
>A<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8),nrow=4)
>A
>B<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8),nrow=3)
>B
>A%*%B
```

```
> A<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8),nrow=4)
> A
     [,1] [,2] [,3]
[1,]    2    0    7
[2,]    5    6    4
[3,]    6    5    9
[4,]    1    8    8
> B<-matrix(c(2,5,6,1,0,6,5,8,7,4,9,8),nrow=3)
> B
     [,1] [,2] [,3] [,4]
[1,]    2    1    5    4
[2,]    5    0    8    9
[3,]    6    6    7    8
> A%*%B
     [,1] [,2] [,3] [,4]
[1,]   46   44   59   64
[2,]   64   29  101  106
[3,]   91   60  133  141
[4,]   90   49  125  140
```

La diagonal:

```
>diag(A)
```

```
> diag(A)
[1] 2 6 9
```

Además, existen muchas otras funciones de uso más general, que también pueden ser utilizadas con matrices o que de hecho resultan particularmente interesantes cuando se aplican sobre matrices u objetos similares. Algunas de ellas, para aplicarlas a matrices, debemos indicar la dimensión en la que queremos que opere (filas o columnas), o seleccionar sólo una fila o columna, operando sobre ella como si se tratase de un vector.

Todas las funciones de cálculos de vectores se pueden aplicar a una matriz si especificamos una dimensión concreta:

```
>A<-matrix(c(5,2,6,0,8,7,6,10,1,0,4,3),nrow=3)
>A
>mean(A[1,])
>sum(A[,4])
```

```
> A<-matrix(c(5,2,6,0,8,7,6,10,1,0,4,3),nrow=3)
> A
     [,1] [,2] [,3] [,4]
[1,]    5    0    6    0
[2,]    2    8   10    4
[3,]    6    7    1    3
> mean(A[1,])
[1] 2.75
> sum(A[,4])
[1] 7
```

Además hay funciones específicas para resumir cálculos por filas o columnas:

```
>rowMeans(A)
>colMeans(A)
```

```
> rowMeans(A)
[1] 2.75 6.00 4.25
> colMeans(A)
[1] 4.333333 5.000000 5.666667 2.333333
```

Existe una función llamada «apply» que permite hacer un bucle sobre una matriz, calculando los valores de una determinada función por filas o por columnas. Más adelante veremos versiones de esta función para listas, tablas, etc. («lapply», «tapply»). La función «apply» es específica de matrices o *data frames*. Para aplicarla, es necesario especificar los siguientes argumentos: nombre del objeto (matriz) sobre el que se aplica, dimensión sobre la que se aplica (con «1» designamos filas, con «2» las columnas), y la función que queremos aplicar. Veamos algún ejemplo.

Con la matriz «A» del ejemplo anterior, queremos que haga las medias de todas las columnas:

```
>apply(A,2,mean)
```

```
> apply(A,2,mean)
[1] 4.333333 5.000000 5.666667 2.333333
```

Como vemos, los resultados son los mismos que con la función «colmeans», ya que le hemos especificado que haga las medias por columnas «2». Esta función y las de su familia se utilizan mucho, porque permiten ahorrarse un bucle.

Otra función interesante es «sweep», que se usa para saber cuánto se desvía los valores de una matriz respecto de un valor determinado. Veamos un ejemplo, con la matriz anterior. En primer lugar se crea un objeto con los valores sobre los que queremos conocer esta desviación. Ahora queremos saber cuánto se desvían respecto de esta suma total de filas, cada elemento de «A», cada uno respecto a la suma de su fila. Debemos utilizar la función «sweep» e indicar que el cálculo lo haga en la dimensión correcta, en este caso filas:

```
>suma<-apply(A,1,sum)
>suma
>sweep(A,1,suma)
```

```
> suma<-apply(A,1,sum)
> suma
[1] 11 24 17
> sweep(A,1,suma)
     [,1] [,2] [,3] [,4]
[1,]   -6  -11   -5  -11
[2,]  -22  -16  -14  -20
[3,]  -11  -10  -16  -14
```

Existen otras funciones importantes que se usan típicamente con datos en formato tipo matriz, como «tapply», «aggregate», «subset», etc. No obstante, las veremos más adelante, al hablar de *data frames* (apartado AII.4.11) ya que resulta más habitual emplearlas en ese caso.

AII.4.10. LISTAS

Una lista en R es simplemente un conjunto de objetos de cualquier tipo. Un gran número de test que realizaremos en R nos devuelven los resultados en formato lista.

AII.4.10.1. Creación de Listas

Se crean con el comando «list», seguido, entre paréntesis, de los nombres de todos aquellos objetos que queramos incluir en la lista. Por ejemplo, si queremos crear una lista incluyendo los objetos «A» (matriz), «v1» (vector) y «x» (número):

```
>A<-matrix(c(5,2,6,0,8,7,6,10,1,0,4,3),nrow=3)
```

```
>A
>v1<-letters[1:10]
>v1
>x<-5
>x
>milista<-list(A,v1,x)
>milista
```

```
> A<-matrix(c(5,2,6,0,8,7,6,10,1,0,4,3),nrow=3)
> A
     [,1] [,2] [,3] [,4]
[1,]    5    0    6    0
[2,]    2    8   10    4
[3,]    6    7    1    3
> v1<-letters[1:10]
> v1
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j"
> x<-5
> x
[1] 5
> milista<-list(A,v1,x)
> milista
[[1]]
     [,1] [,2] [,3] [,4]
[1,]    5    0    6    0
[2,]    2    8   10    4
[3,]    6    7    1    3
```

AII.4.10.2. Trabajando con listas

En el caso de las listas, se accede a sus componentes a través de los dobles corchetes «[[]]», indicando el número de componente que queremos. En el caso de la lista anterior, si nos interesa acceder al elemento 2:

```
>milista[[2]]
```

```
> milista[[2]]
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j"
```

Si los componentes de la lista tienen un nombre, es posible acceder a ellos con el doble corchete o con el operador «$». Primero nombremos a los componentes de «milista»:

```
>names(milista)<-c("A","v1","x")
>milista
```

```
> names(milista)<-c("A","v1","x")
> milista
$A
     [,1] [,2] [,3] [,4]
[1,]    5    0    6    0
[2,]    2    8   10    4
[3,]    6    7    1    3

$v1
 [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j"

$x
[1] 5
```

Ahora accedamos a cada uno de ellos por su nombre con doble corchete o con el operador «$»:

```
>milista[["x"]]
>milista$A
```

```
> milista[["x"]]
[1] 5
> milista$A
     [,1] [,2] [,3] [,4]
[1,]    5    0    6    0
[2,]    2    8   10    4
[3,]    6    7    1    3
```

Sobre los componentes de una lista es posible realizar cualquier tipo de operación, o aplicar cualquier función que sea viable para ese tipo de objeto. Simplemente tendríamos que indicar el elemento de la lista sobre el que queremos operar. Si por ejemplo, en la lista anterior, quisiéramos extraer la diagonal de la matriz (lógicamente, esto sólo se podría hacer con el elemento A):

```
>diag(milista$A)
```

```
> diag(milista$A)
[1] 5 8 1
```

o también:

```
> diag(milista[["A"]])
```

```
> diag(milista[["A"]])
[1] 5 8 1
```

Pero en algún caso nos podría interesar aplicar una sola función a todos los componentes de una lista. En este caso, podemos utilizar las funciones «lapply» o sapply» (esta última la hemos visto para vectores, pero funciona también para listas), donde «lapply» se puede considerar la versión de «apply» para listas.

En el caso de la lista que creamos «milista», puesto que sus componentes son muy distintos, existirán pocas funciones que se puedan aplicar a todos los componentes a la vez y nos devuelvan un resultado. Sería posible, por ejemplo, con la función «max», que nos devuelve el valor máximo de cada componente:

>lapply(milista, max)

```
> lapply(milista, max)
$A
[1] 10

$v1
[1] "j"

$x
[1] 5
```

Obviamente, para el vector «v1», el valor máximo utiliza el orden alfabético. Veamos que con «sapply», el resultado es el mismo, son dos funciones muy parecidas, pero esta última devuelve un formato de resultados similar a como si hubiese operado sobre un vector en lugar de una lista:

>sapply(milista, max)

```
> sapply(milista, max)
   A   v1    x
"10"  "j"  "5"
```

AII.4.11. DATA FRAMES

Un *data frame* es un conjunto de datos, en formato similar a matriz, en el cual tenemos los valores de una serie de variables. Habitualmente, cada variable constituye una columna del *data frame*, y cada caso, una fila. Son elementos de especial importancia, ya que son las estructuras más habituales de los datos estadísticos, y la base sobre la que se suelen hacer cálculos y tests.

AII.4.11.1. Creación de *data frames*

El comando «data.frame()», convierte en *data frame* cualquier objeto o conjunto de objetos que tengan las dimensiones adecuadas (si, por ejemplo, tenemos varios vectores, éstos deben tener la misma longitud). Puede ser especialmente útil en este caso el comando «as.data.frame()», que nos permitiría convertir un objeto en *data frame*, si fuese posible. Veamos varios ejemplos.

En primer lugar crearemos un *data frame* con tres variables y 25 casos a partir de tres vectores que definiremos previamente:

```
>v1<-1:25
>v1
>v2<-sample(v1)
>v2
```

```
>v3<-gl(5,5)
>v3
```

```
> v1<-1:25
> v1
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
> v2<-sample(v1)
> v2
 [1]  5 14  2 21 13  1  4 19  7 23 12  8  6 16 24 20 11 15 18  9 10 22 25  3 17
> v3<-gl(5,5)
> v3
 [1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 5 5 5 5 5
Levels: 1 2 3 4 5
```

Así pues tenemos el vector «v1», que es una secuencia de enteros de 1 a 25, el vector «v2», que son los elementos de la secuencia anterior en orden aleatorio, y el vector «v3», que es un factor con cinco niveles, y cinco valores para cada nivel. Como todos tienen longitud 25, podemos unirlos en un *data frame*:

```
>misdatos<-data.frame(v1,v2,v3)
>misdatos
```

```
> misdatos<-data.frame(v1,v2,v3)
> misdatos
   v1 v2 v3
1   1  5  1
2   2 14  1
3   3  2  1
4   4 21  1
5   5 13  1
6   6  1  2
7   7  4  2
8   8 19  2
9   9  7  2
10 10 23  2
11 11 12  3
12 12  8  3
13 13  6  3
14 14 16  3
```

El *data frame* tiene por nombres de variables los de los antiguos vectores. Los casos se numeran en una columna situada a la izquierda.

En R es posible modificar clases de objetos, como ya hemos visto, por medio de funciones del tipo «*as.nombre de la clase()*». De esta forma podríamos crear un *data frame*, convirtiendo directamente un vector, o una matriz, como en el siguiente ejemplo. Si tenemos la matriz «A», la podemos convertir directamente en *data frame*:

```
>A<-matrix(c(seq(1:50)),nrow=10)
>A
>datosA<-as.data.frame(A)
>datosA
```

```
> A<-matrix(c(seq(1:50)),nrow=10)
> A
      [,1] [,2] [,3] [,4] [,5]
 [1,]    1   11   21   31   41
 [2,]    2   12   22   32   42
 [3,]    3   13   23   33   43
 [4,]    4   14   24   34   44
 [5,]    5   15   25   35   45
 [6,]    6   16   26   36   46
 [7,]    7   17   27   37   47
 [8,]    8   18   28   38   48
 [9,]    9   19   29   39   49
[10,]   10   20   30   40   50
> datosA<-as.data.frame(A)
> datosA
   V1 V2 V3 V4 V5
1   1 11 21 31 41
2   2 12 22 32 42
3   3 13 23 33 43
4   4 14 24 34 44
5   5 15 25 35 45
6   6 16 26 36 46
7   7 17 27 37 47
8   8 18 28 38 48
9   9 19 29 39 49
10 10 20 30 40 50
```

Vemos que el *data frame* creado contiene los mismos datos que la matriz, y asume que las columnas son variables y las filas casos. Ante la falta de nombres para las variables, por defecto las nombra sucesivamente como V1… V*n*, siendo *n* el número de columnas.

Como se mencionó en apartados anteriores, existen ejemplos con datos dentro de muchos paquetes de R. La mayoría son de clase *data frame*. Por ejemplo, ya habíamos hablado del conjunto de datos «iris», que se encuentra en el paquete «datasets», el cual se carga por defecto cuando se instala R.

```
>data(iris)
>class(iris)
[1] "data.frame"
```

Una vez que disponemos de un *data frame* cargado en nuestra sesión, la función «summary» nos provee un resumen con información sobre el conjunto de datos: nos da los valores máximos, mínimos, medias, medianas, cuartiles y el *n* de cada variable. Podemos ver un ejemplo con el conjunto de datos «iris»:

```
>data(iris)
>summary(iris)
```

```
> data(iris)
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

Cuando creamos un *data frame*, como en el ejemplo de «misdatos», es posible que nos interese cambiar los nombres de las variables. Hay dos formas bastante comunes de hacerlo. En primer lugar, nombrándolas directamente al crear el *data frame*. Volviendo a «misdatos», llamaremos a las tres variables «x», «y», y «factor»:

```
>misdatos<-data.frame(x=v1,y=v2,factor=v3)
>misdatos
```

```
> misdatos<-data.frame(x=v1,y=v2,factor=v3)
> misdatos
   x  y factor
1  1  7      1
2  2 10      1
3  3 22      1
4  4  5      1
5  5 14      1
```

Otra alternativa sería utilizar la función «names»:

```
>names(misdatos)<-c("x","y","factor")
>misdatos
```

```
> names(misdatos)<-c("x","y","factor")
> misdatos
   x  y factor
1  1  7      1
2  2 10      1
3  3 22      1
4  4  5      1
5  5 14      1
```

AII.4.11.2. Seleccionando variables y casos en *data frames*

Entre las operaciones más importantes que podemos realizar con *data frames*, está la de seleccionar elementos dentro de estos (filas, columnas, subconjuntos)

por medio de índices o con condiciones lógicas. Una a una iremos viendo diversas posibilidades. Muchas de ellas son similares a las que hemos visto en el apartado de matrices. El operador que se utiliza para localizar elementos dentro del *data frame* son los corchetes simples, al igual que en vectores y matrices. Para los ejemplos usaremos el conjunto de datos «iris»:

```
>data(iris)
```

Si queremos seleccionar una fila, utilizando números, debemos indicar entre corchetes el número de fila, a la izquierda de una coma. El espacio vacío a la derecha de la coma indica que seleccionamos todas las columnas:

```
>iris[20,]
```

```
> iris[20,]
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
20          5.1         3.8          1.5         0.3  setosa
```

Es interesante destacar que se conserva el número de fila antiguo (al contrario que en el caso de las matrices). Ahora, si quisiéramos seleccionar una columna:

```
>iris[,3]
```

```
> iris[,3]
  [1] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 1.5 1.6 1.4 1.1 1.2 1.5
 [19] 1.7 1.5 1.7 1.5 1.0 1.7 1.9 1.6 1.6 1.5 1.4 1.6 1.6 1.5 1.5 1.4
 [37] 1.3 1.4 1.3 1.5 1.3 1.3 1.3 1.6 1.9 1.4 1.6 1.4 1.5 1.4 4.7 4.5
 [55] 4.6 4.5 4.7 3.3 4.6 3.9 3.5 4.2 4.0 4.7 3.6 4.4 4.5 4.1 4.5 3.9
 [73] 4.9 4.7 4.3 4.4 4.8 5.0 4.5 3.5 3.8 3.7 3.9 5.1 4.5 4.5 4.7 4.4
 [91] 4.4 4.6 4.0 3.3 4.2 4.2 4.2 4.3 3.0 4.1 6.0 5.1 5.9 5.6 5.8 6.6
[109] 5.8 6.1 5.1 5.3 5.5 5.0 5.1 5.3 5.5 6.7 6.9 5.0 5.7 4.9 6.7 4.9
[127] 4.8 4.9 5.6 5.8 6.1 6.4 5.6 5.1 5.6 6.1 5.6 5.5 4.8 5.4 5.6 5.1
```

Si nos interesa seleccionar un dato concreto, podemos indicar fila y columna:

```
>iris[20,3]
```

```
> iris[20,3]
[1] 1.5
```

También es posible seleccionar varias filas, por ejemplo, salteadas:

```
>iris[c(15,29,23),]
```

```
> iris[c(15,29,23),]
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
15          5.8         4.0          1.2         0.2  setosa
29          5.2         3.4          1.4         0.2  setosa
23          4.6         3.6          1.0         0.2  setosa
```

Y lo mismo para columnas:

>iris[,c(1,3,4)]

```
> iris[,c(1,3,4)]
    Sepal.Length Petal.Length Petal.Width
1            5.1          1.4         0.2
2            4.9          1.4         0.2
3            4.7          1.3         0.2
4            4.6          1.5         0.2
5            5.0          1.4         0.2
```

Y a la vez filas y columnas salteadas:

>iris[c(15,20,23),c(1,3,4)]

```
> iris[c(15,20,23),c(1,3,4)]
   Sepal.Length Petal.Length Petal.Width
15          5.8          1.2         0.2
20          5.1          1.5         0.3
23          4.6          1.0         0.2
```

También es posible seleccionar un conjunto de filas y/o columnas contiguas con el operador de secuencias. Veamos un ejemplo en el que seleccionamos a la vez filas y columnas:

>iris[3:14,2:4]

```
> iris[3:14,2:4]
   Sepal.Width Petal.Length Petal.Width
3          3.2          1.3         0.2
4          3.1          1.5         0.2
5          3.6          1.4         0.2
6          3.9          1.7         0.4
7          3.4          1.4         0.3
8          3.4          1.5         0.2
9          2.9          1.4         0.2
10         3.1          1.5         0.1
11         3.7          1.5         0.2
12         3.4          1.6         0.2
```

Todos estos tipos de índices son combinables entre sí, para hacer cualquier tipo de selección.

Una de las formas más interesantes de seleccionar datos dentro de *data frames* consiste en utilizar expresiones lógicas. Se utiliza para seleccionar casos (filas) ya que con variables carece de sentido. Veamos una serie de ejemplos.

En primer lugar, de ahora en adelante, utilizaremos la función *attach()*, que nos permite adjuntar en la memoria de forma permanente un conjunto de datos. Así, dicho conjunto será utilizado como el entorno en el que R busca preferentemente los nombres de objetos. De esta manera podemos escribir los nombres de las variables tal cual, sin utilizar el operador de localización «$», y podremos nombrar a las variables por su nombre directamente:

```
> attach(iris)
> Petal.Length
```

```
> attach(iris)
> Petal.Length
  [1] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 1.5 1.6 1.4 1.1 1.2 1.5 1.3 1.4 1.7 1.5
 [21] 1.7 1.5 1.0 1.7 1.9 1.6 1.6 1.5 1.4 1.6 1.6 1.5 1.5 1.4 1.5 1.2 1.3 1.4 1.3 1.5
 [41] 1.3 1.3 1.3 1.6 1.9 1.4 1.6 1.4 1.5 1.4 4.7 4.5 4.9 4.0 4.6 4.5 4.7 3.3 4.6 3.9
 [61] 3.5 4.2 4.0 4.7 3.6 4.4 4.5 4.1 4.5 3.9 4.8 4.0 4.9 4.7 4.3 4.4 4.8 5.0 4.5 3.5
 [81] 3.8 3.7 3.9 5.1 4.5 4.5 4.7 4.4 4.1 4.0 4.4 4.6 4.0 3.3 4.2 4.2 4.2 4.3 3.0 4.1
[101] 6.0 5.1 5.9 5.6 5.8 6.6 4.5 6.3 5.8 6.1 5.1 5.3 5.5 5.0 5.1 5.3 5.5 6.7 6.9 5.0
[121] 5.7 4.9 6.7 4.9 5.7 6.0 4.8 4.9 5.6 5.8 6.1 6.4 5.6 5.1 5.6 6.1 5.6 5.5 4.8 5.4
[141] 5.6 5.1 5.1 5.9 5.7 5.2 5.0 5.2 5.4 5.1
```

en lugar de tener que escribir:

```
> iris$Petal.Length
```

```
> iris$Petal.Length
  [1] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 1.5 1.6 1.4 1.1
 [23] 1.0 1.7 1.9 1.6 1.6 1.5 1.4 1.6 1.6 1.5 1.5 1.4 1.5 1.2
 [45] 1.9 1.4 1.6 1.4 1.5 1.4 4.7 4.5 4.9 4.0 4.6 4.5 4.7 3.3
 [67] 4.5 4.1 4.5 3.9 4.8 4.0 4.9 4.7 4.3 4.4 4.8 5.0 4.5 3.5
 [89] 4.1 4.0 4.4 4.6 4.0 3.3 4.2 4.2 4.2 4.3 3.0 4.1 6.0 5.1
[111] 5.1 5.3 5.5 5.0 5.1 5.3 5.5 6.7 6.9 5.0 5.7 4.9 6.7 4.9
[133] 5.6 5.1 5.6 6.1 5.6 5.5 4.8 5.4 5.6 5.1 5.1 5.9 5.7 5.2
```

En el caso de que se quiera seleccionar casos (filas) que tengan un valor determinado de una variable, por ejemplo, una anchura de sépalo de 3:

```
>iris[Sepal.Width==3.0,]
```

```
> iris[Sepal.Width==3.0,]
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
2            4.9           3          1.4         0.2     setosa
13           4.8           3          1.4         0.1     setosa
14           4.3           3          1.1         0.1     setosa
26           5.0           3          1.6         0.2     setosa
39           4.4           3          1.3         0.2     setosa
46           4.8           3          1.4         0.3     setosa
62           5.9           3          4.2         1.5 versicolor
67           5.6           3          4.5         1.5 versicolor
```

En la ventana anterior se observa cómo selecciona sólo aquellas filas en las que dicha variable toma valor 3.

Es posible complicar un poco más la condición lógica. Por ejemplo, seleccionaremos aquellos casos (filas) en los que la variable anchura de sépalo tenga valor 3 y la longitud de pétalo 4,5:

```
>iris[(Sepal.Width==3.0)&(Petal.Length==4.5),]
```

```
> iris[(Sepal.Width==3.0)&(Petal.Length==4.5),]
   Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
67          5.6           3          4.5         1.5 versicolor
85          5.4           3          4.5         1.5 versicolor
```

Cualquier otra combinación de operadores lógicos que se nos ocurra resultaría igualmente válida. Otro caso interesante sería, por ejemplo, seleccionar casos que tengan valores de las variables seleccionadas dentro de un rango determinado:

>iris[(Sepal.Width<3.5)&(Petal.Length>4),]

```
> iris[(Sepal.Width<3.5)&(Petal.Length>4),]
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
51           7.0         3.2          4.7         1.4 versicolor
52           6.4         3.2          4.5         1.5 versicolor
53           6.9         3.1          4.9         1.5 versicolor
55           6.5         2.8          4.6         1.5 versicolor
```

Hasta ahora hemos visto selección de casos en variables numéricas, pero también es interesante el caso de las variables categóricas. Si, por ejemplo, nos interesa seleccionar los datos de «iris» referidos tan sólo a una de las especies, «virginica» (el valor de la variable categórica debe entrecomillarse):

>iris[Species==”virginica”,]

```
> iris[Species=="virginica",]
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
101          6.3         3.3          6.0         2.5 virginica
102          5.8         2.7          5.1         1.9 virginica
103          7.1         3.0          5.9         2.1 virginica
104          6.3         2.9          5.6         1.8 virginica
105          6.5         3.0          5.8         2.2 virginica
106          7.6         3.0          6.6         2.1 virginica
```

O, si queremos seleccionar dos de las tres especies, podemos indicar que se seleccionen los casos distintos a la especie que queremos dejar fuera:

>iris[Species!=”versicolor”,]

```
> iris[Species!="versicolor",]
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
1            5.1         3.5          1.4         0.2    setosa
2            4.9         3.0          1.4         0.2    setosa
3            4.7         3.2          1.3         0.2    setosa
4            4.6         3.1          1.5         0.2    setosa
5            5.0         3.6          1.4         0.2    setosa
6            5.4         3.9          1.7         0.4    setosa
7            4.6         3.4          1.4         0.3    setosa
8            5.0         3.4          1.5         0.2    setosa
9            4.4         2.9          1.4         0.2    setosa
```

Al igual que en vectores y matrices, es posible eliminar componentes de un *data frame* utilizando símbolos negativos, como por ejemplo en este caso que eliminamos las filas 10 a 20.

>iris[-(10:20),]

```
> iris[-(10:20),]
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
21           5.4         3.4          1.7         0.2     setosa
22           5.1         3.7          1.5         0.4     setosa
23           4.6         3.6          1.0         0.2     setosa
```

Son posibles todas las combinaciones imaginables de los recursos descritos hasta ahora para seleccionar casos en *data frames*. Además, es posible insertar funciones para establecer las condiciones a la hora de seleccionar. En este ejemplo, queremos seleccionar los casos de la variable longitud de sépalo que sean superiores a su valor promedio. Entonces podemos insertar la función que calcula el valor promedio dentro de la propia condición definida para seleccionar. Esto es posible hacerlo con cualquier otra función.

>iris[Sepal.Length>mean(Sepal.length),]

```
> iris[Sepal.Length>mean(Sepal.Length),]
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
51           7.0         3.2          4.7         1.4 versicolor
52           6.4         3.2          4.5         1.5 versicolor
53           6.9         3.1          4.9         1.5 versicolor
55           6.5         2.8          4.6         1.5 versicolor
57           6.3         3.3          4.7         1.6 versicolor
59           6.6         2.9          4.6         1.3 versicolor
62           5.9         3.0          4.2         1.5 versicolor
63           6.0         2.2          4.0         1.0 versicolor
```

Existen algunas funciones específicas para seleccionar casos, como «subset» y «which», que veremos en el siguiente apartado, que trata sobre el uso de funciones en *data frames.*

AII.4.11.3. Funciones importantes en *data frames*

Dentro de las muchas funciones que es posible utilizar para operar con *data frames*, mostraremos algunos ejemplos de aquellas que nos parecen más relevantes para usuarios que deben iniciarse en el manejo de R.

En relación con el apartado anterior, existen dos funciones interesantes para seleccionar datos por casos. Una de ellas es «which», que se emplea, de forma muy intuitiva, para seleccionar aquellos datos que cumplan una condición específica. Continuando con los datos «iris» que se vio cómo cargarlos en el apartado anterior:

>iris[which(Sepal.Length>7),]

```
> iris[which(Sepal.Length>7),]
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
103          7.1         3.0          5.9         2.1 virginica
106          7.6         3.0          6.6         2.1 virginica
108          7.3         2.9          6.3         1.8 virginica
110          7.2         3.6          6.1         2.5 virginica
118          7.7         3.8          6.7         2.2 virginica
```

>iris[which(Species==”setosa”),]

```
> iris[which(Species=="setosa"),]
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           5.1         3.5          1.4         0.2  setosa
2           4.9         3.0          1.4         0.2  setosa
3           4.7         3.2          1.3         0.2  setosa
```

La función «subset» resulta muy útil para extraer secciones de *data frames*. Por ejemplo, para seleccionar diferentes valores de variables categóricas se usa indicando cuál es la variable sobre la que vamos a operar y después se añade «%in%» antes de un vector que contiene las categorías que queremos seleccionar:

>subset(iris,Species %in% c(“virginica”,”versicolor”))

```
> subset(iris,Species %in% c("virginica","versicolor"))
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
51           7.0         3.2          4.7         1.4 versicolor
52           6.4         3.2          4.5         1.5 versicolor
53           6.9         3.1          4.9         1.5 versicolor
```

Otra forma diferente de seleccionar categorías (tanto casos como variables) es:

> iris1 <- iris[iris$Species %in% c("virginica","versicolor"), c("Sepal.Length","Species")]
> iris1

```
> iris1 <- iris[iris$Species %in% c("virginica","versicolor"), c("Sepal.Length","Species")]
> iris1
    Sepal.Length    Species
51           7.0 versicolor
52           6.4 versicolor
53           6.9 versicolor
```

La instrucción «subset» también se puede usar para extraer variables completas, como seleccionando columnas:

>subset(iris,select=c(“Sepal.Length”,”Species”))

```
> subset(iris,select=c("Sepal.Length","Species"))
    Sepal.Length    Species
1            5.1     setosa
2            4.9     setosa
3            4.7     setosa
```

Otras funciones de importante uso en los *data frames* son las que nos permiten gestionar datos perdidos. La función «na.omit()» elimina temporalmente todos los casos que contengan al menos un valor perdido. Se usa de la siguiente forma, con el ejemplo de *iris*:

```
>na.omit(iris)
```

Otra función similar, y que se usa de la misma forma es «na.exclude()». Se diferencia de «na.omit()» en que, aunque no tiene en cuenta los valores perdidos, no elimina por completo los casos que tengan valores perdidos. Por ejemplo, si ajustamos un modelo a unos datos, y extraemos los residuos, «na.exclude()» no tendría en cuenta valores perdidos para los cálculos del modelo, pero sí para el número de residuos generado, que sería igual al total de datos, aunque los residuos correspondientes a los casos con valores perdidos serían a su vez valores perdidos ("NA"), ya que no pueden ser calculados. Con «na.omit()» el número de residuos sería igual al total de datos menos el número de casos con valores perdidos.

```
>na.exclude(iris)
```

Otras funciones interesantes sirven para aumentar las dimensiones de los *data frames*: añadir filas, columnas o unir diferentes *data frames*. Por ejemplo, «rbind()» y «cbind()» se usan igual que con matrices, y sirven para añadir filas y columnas, respectivamente. Las dimensiones del *data frame* y de la fila o columna añadida deben ser compatibles. Si creamos un vector con el mismo número de casos que «iris», podemos añadirlo como columna:

```
>v1<-1:150
>cbind(iris,v1)
```

```
> v1<-1:150
> cbind(iris,v1)
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species  v1
1            5.1         3.5          1.4         0.2     setosa   1
2            4.9         3.0          1.4         0.2     setosa   2
3            4.7         3.2          1.3         0.2     setosa   3
4            4.6         3.1          1.5         0.2     setosa   4
5            5.0         3.6          1.4         0.2     setosa   5
6            5.4         3.9          1.7         0.4     setosa   6
7            4.6         3.4          1.4         0.3     setosa   7
8            5.0         3.4          1.5         0.2     setosa   8
```

La función «merge()» sirve para unir dos *data frames* con el mismo número de casos. Se usa del siguiente modo:

>merge(*data frame1,data frame2*)

La función «apply» ya la hemos visto en el apartado de matrices. Para *data frames* funcionaría de la misma manera. En «iris» debemos usarla sólo en las columnas de la 1 a la 4, porque la 5 es de una variable categórica:

> apply(iris[,1:4],2,mean)

```
> apply(iris[,1:4],2,mean)
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
    5.843333     3.057333     3.758000     1.199333
```

Otro tipo de funciones interesantes son aquellas que nos permiten, de diversas maneras, obtener resultados de *data frames* por grupos. Una de la misma familia de «apply» es «tapply», que nos permite obtener los valores de una función para una variable en diversos grupos. Con «tapply» debe indicarse entre paréntesis la variable sobre la que queremos calcular la función, la variable que conforma los grupos, y por último, la función. En este ejemplo, queremos saber cuál es la mediana de longitud de sépalo por especie, dentro de «iris»:

>tapply(Sepal.Length,Species,median)

```
> tapply(Sepal.Length,Species,median)
    setosa versicolor  virginica
       5.0        5.9        6.5
```

La función «aggregate» opera de un modo muy similar, pero indicando con el argumento «by» una lista que debe contener las categorías sobre las que queremos calcular los valores de la función, y a las que además le podemos poner un nombre. El formato de salida es distinto:

>aggregate(Sepal.Length,by=list(Especies=Species),median)

```
> aggregate(Sepal.Length,by=list(Especies=Species),median)
    Especies   x
1     setosa 5.0
2 versicolor 5.9
3  virginica 6.5
```

Otra función muy interesante relacionada con las anteriores es «by». Sus argumentos son: el nombre del *data frame* sobre el que se aplica, la variable de agrupación de casos, que es un factor con diferentes niveles, y la función (por lo que es muy parecida a «tapply»). Aquí tenemos un ejemplo en el que queremos aplicar la función que calcula el coeficiente de correlación («cor») a dos de las variables del *data frame* «iris», y que lo segmente por especies. Este ejemplo sería mucho más complicado con «tapply» y «aggregate», puesto que es una función que usa dos dimensiones:

>by(iris[,c("Sepal.Length","Petal.Length")],Species,cor)

```
> by(iris[,c("Sepal.Length","Petal.Length")],Species,cor)
Species: setosa
             Sepal.Length Petal.Length
Sepal.Length    1.0000000    0.2671758
Petal.Length    0.2671758    1.0000000
------------------------------------------------------------
Species: versicolor
             Sepal.Length Petal.Length
Sepal.Length     1.000000     0.754049
Petal.Length     0.754049     1.000000
```

AII.4.12. EDITOR DE DATOS

R posee un editor de datos que nos permite visualizar cualquier *data frame* que hayamos cargado. Se puede acceder a este editor a través de menús (según la GUi con la que estemos trabajando) o bien con las funciones edit() o fix(). Utilizando iris como ejemplo:

```
>edit(iris)
>fix(iris)
```

Este es el aspecto del editor de datos:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	var6
1	5.1	3.5	1.4	0.2	setosa	
2	4.9	3	1.4	0.2	setosa	
3	4.7	3.2	1.3	0.2	setosa	
4	4.6	3.1	1.5	0.2	setosa	
5	5	3.6	1.4	0.2	setosa	
6	5.4	3.9	1.7	0.4	setosa	
7	4.6	3.4	1.4	0.3	setosa	
8	5	3.4	1.5	0.2	setosa	
9	4.4	2.9	1.4	0.2	setosa	
10	4.9	3.1	1.5	0.1	setosa	
11	5.4	3.7	1.5	0.2	setosa	

Las capacidades de edición son muy limitadas, se reducen a copiar, pegar, cambiar el contenido de las celdas o añadir valores a alguna nueva variable, si existe, como en este caso «var6».

AII.4.13. TEXTO

Además de existir algún tipo de editor de texto en las GUis, dentro de R tenemos funciones que sirven para escribir texto en la propia consola. La función «print» imprime objetos o caracteres (tienen que ir entre comillas):

```
>X<-matrix(c(1,0,1,0,1,1,1,0,1,0,0,1,0,1,0),nrow=3)
>print("X es una matriz")
```

```
> X<-matrix(c(1,0,1,0,1,1,1,0,1,0,0,1,0,1,0),nrow=3)
> print(X)
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    0    1    0    0
[2,]    0    1    0    0    1
[3,]    1    1    1    1    0
> print("X es una matriz")
[1] "X es una matriz"
```

Una función muy parecida es «cat», pero devuelve los objetos o caracteres pegados en un formato más simple, encadenándolo con el siguiente comando:

```
>cat(X)
>cat("X es una matriz")
```

```
> cat(X)
1 0 1 0 1 1 1 0 1 0 0 1 0 1 0> cat("X es una matriz")
X es una matriz>
```

Es posible formatear el texto que se suministra a estas funciones de manera que sea un poco más elaborado. La función «paste» sirve para concatenar objetos, variables y texto. Un argumento muy importante que posee esta función es el separador «sep» que se emplea para concatenar cada elemento. Puede ser un espacio en blanco «" "», coma «",»», guión bajo «"_"» etc. En este ejemplo utilizamos la función «paste» para concatenar elementos en una serie de objetos hasta que llegamos a uno definitivo llamado «formula», que puede ser la fórmula de diferentes tipos de modelo, con la particularidad de que para incluir o eliminar variables independientes en el modelo, sólo tendríamos que modificar los elementos del vector »variables» y, automáticamente, la fórmula cambia allá donde se esté aplicando (ya veremos a la hora de hacer pruebas estadísticas cómo es el formato de las fórmulas para modelos en R).

Primero, tenemos todas las variables en un *data frame* con sus nombres, y luego vamos creando una serie de objetos hasta llegar a uno con formato de fórmula, que se usa directamente en un modelo lineal (función «lm»). Para este ejemplo, vamos a utilizar el conjunto de datos *airquality*, presente en el paquete *datasets*, que como sabemos, está cargado por defecto:

```
>data(airquality)
>attach(airquality)
>variables<-paste("Solar.R","Wind","Temp","Month","Day",sep="+")
>variables
>formula<-as.formula(paste("Ozone~",variables,sep=""))
>formula
>modelo<-lm(formula)
>summary(modelo)
```

```
> data(airquality)
> attach(airquality)
> variables<-paste("Solar.R","Wind","Temp","Month","Day",sep="+")
> variables
[1] "Solar.R+Wind+Temp+Month+Day"
> formula<-as.formula(paste("Ozone~",variables,sep=""))
> formula
Ozone ~ Solar.R + Wind + Temp + Month + Day
> modelo<-lm(formula)
> summary(modelo)

Call:
lm(formula = formula)

Residuals:
    Min      1Q  Median      3Q     Max
-37.014 -12.284  -3.302   8.454  95.348
```

Para formatear texto se utilizan una serie de caracteres, de la siguiente forma: «\n» indica nueva línea, «\b» espacio en blanco detrás, «\t» tabulador, y otros similares. Por ejemplo:

>cat("Variables empleadas en el análisis:\n","\tModelo 1:var1,2,3,4\n","
\tModelo 2: var1,2,3\n")

```
> cat("Variables empleadas en el análisis:\n","\tModelo 1: var1,2,3,4\n","\tModelo 2: var1,2,3\n")
Variables empleadas en el análisis:
        Modelo 1: var1,2,3,4
        Modelo 2: var1,2,3
```

AII.4.14. GRÁFICOS

AII.4.14.1. Aspectos generales

Para crear gráficos existen, por un lado, funciones que generan diferentes tipos de formato de gráfico y, por otro, funciones que pueden crear nuevos elementos dentro de esos gráficos. Entre las que sirven para crear gráficos, dentro de la librería "graphics" (instalada y cargada por defecto) tenemos:

>plot(): función genérica para crear gráficos con objetos de R
>barplot(): gráfico de barras.
>hist(): histograma.
>pie(): gráfico de tarta.
>boxplot(): diagrama de cajas.
>contour(): gráfico de líneas de contorno.
>persp(): gráfico en 3D.
>ts.plot(): serie temporal.
>pairs(): gráficos de dispersión de todos los pares de variables disponibles.
>coplot(): gráficos condicionados.

Para acceder a una lista completa de todas las funciones dentro de esta librería, teclearíamos la siguiente instrucción:

```
>library(help="graphics")
```

Además existen, como hemos visto anteriormente, librerías que contienen funciones gráficas adicionales. Se instalan y cargan como los paquetes, no obstante muchas de ellas ya están instaladas por defecto y solamente necesitamos cargarlas. Estas son algunas: "lattice", "rgl", "vcd", "maps", etc. Las funciones gráficas que posee cada librería pueden consultarse de la misma forma que la descrita en el párrafo anterior, reemplazando el nombre de la librería por la correspondiente en cada caso.

Resultan fundamentales para trabajar con gráficos los argumentos que se le pueden proporcionar a cada una de estas funciones. La mejor forma de consultar la lista de argumentos disponibles es, como siempre, a través de la ayuda, de esta forma:

```
>help(Nombre de la función)
```

Tenemos argumentos para añadir títulos de gráfico, subtítulos, rótulos de ejes, leyendas, diferentes tipos de símbolo, color de símbolo, de líneas, cambiar tipo de fuente, tamaño de fuente, posición de los elementos, etc. Los valores de estos argumentos pueden estar codificados por números, caracteres, o ambos.

Como decíamos, existe otro tipo de funciones que se emplean para modificar los gráficos introduciendo nuevos objetos. Algunas de las más básicas son:

```
>legend(): añade leyenda.
>title(): añade título.
>text(): añade texto.
>lines(): añade líneas.
>points(): añade puntos.
>abline():añade una línea recta, por ejemplo una regresión lineal.
>arrows(): añade flechas.
>par(): reconfigura parámetros de gráficos
```

Se pueden consultar todas las existentes en cada librería utilizando la misma instrucción indicada más arriba:

```
>library(help="nombre de la librería")
```

Como vemos, algunas de estas funciones nos permiten hacer cosas que ya hemos visto dentro de los argumentos de las funciones gráficas. Por ejemplo, el título se puede añadir en el argumento «main», dentro de la función gráfica, o con una nueva función «title()»

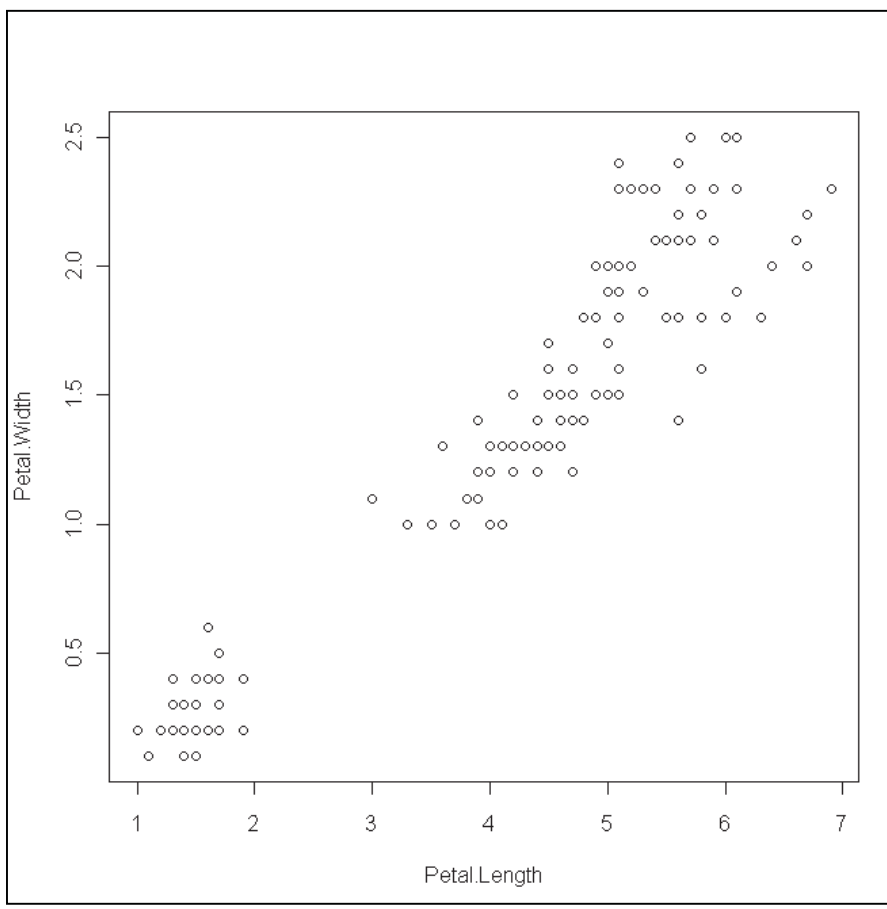
Existen demos que nos muestran buenos ejemplos de gráficos con su correspondiente código:

```
>demo(graphics)
>demo(image)
>demo(persp)
```

AII.4.14.2. Gráfico de dispersión

Para el ejemplo se usará el conjunto de datos *iris*, y se representarán dos variables que deberían presentar una buena relación lineal entre ellas, como «Petal.Length» y «Petal.Width». Para el gráfico de dispersión más simple, utilizamos la función «plot()» sin modificar ninguno de sus argumentos por defecto. Las instrucciones están en el archivo **AII.1.R**.

```
>data(iris)
>attach(iris)
>plot(Petal.Length,Petal.Width)
```



Ahora podemos ir añadiendo y modificando elementos en el gráfico. En primer lugar, le cambiaremos los símbolos. Esto se hace con el argumento «pch», dentro de la función «plot()». Cada tipo de símbolo está representado por un número. En la ayuda se puede consultar a qué tipo de símbolo representa ese número. En nuestro caso, queremos utilizar círculos opacos, a los que les corresponde el número 19:

```
>plot(Petal.Length,Petal.Width,pch=19)
```

Ahora cambiamos el rótulo de los ejes, de forma que el nombre de las variables tenga un mejor formato y se indiquen las unidades. Esto podríamos hacerlo con los argumentos «xlab» e «ylab», dentro de la función «plot()»:

```
>plot(Petal.Length,Petal.Width,pch=19,xlab="Length (cm)",ylab="Width (cm)")
```

Ahora le podemos añadir un título, para lo cual existen dos opciones: o bien el argumento «main», dentro de la función «plot()» o bien otra función llamada «title()». En este caso optaremos por la primera opción.

```
>plot(Petal.Length,Petal.Width,pch=19,xlab="Length (cm)",ylab="Width
(cm)",main="Petal length vs Petal width")
```

Ahora haremos una serie de ajustes en el tipo de letra: convertiremos en negrita la fuente de la escala en los ejes (con «font.axis»), la etiqueta de los ejes (con «font.lab»), aumentaremos el tamaño de fuente en el título (con «cex.main») y en la etiqueta de los ejes (con «cex.lab»). Además, cambiaremos la fuente en todo el gráfico para que sea Times New Roman (con «family»). Los valores de cada uno de los argumentos anteriores serían los siguientes: para negrita, el valor es 2, para el tamaño de fuente, se usa un número que es una proporción relativa al tamaño actual (1). Así pues, si lo cambiamos a 0,5, el tamaño se reduciría un 50%, si lo aumentamos a 1,5, se aumentaría un 50%. El valor de «family» para Times New Roman es serif.

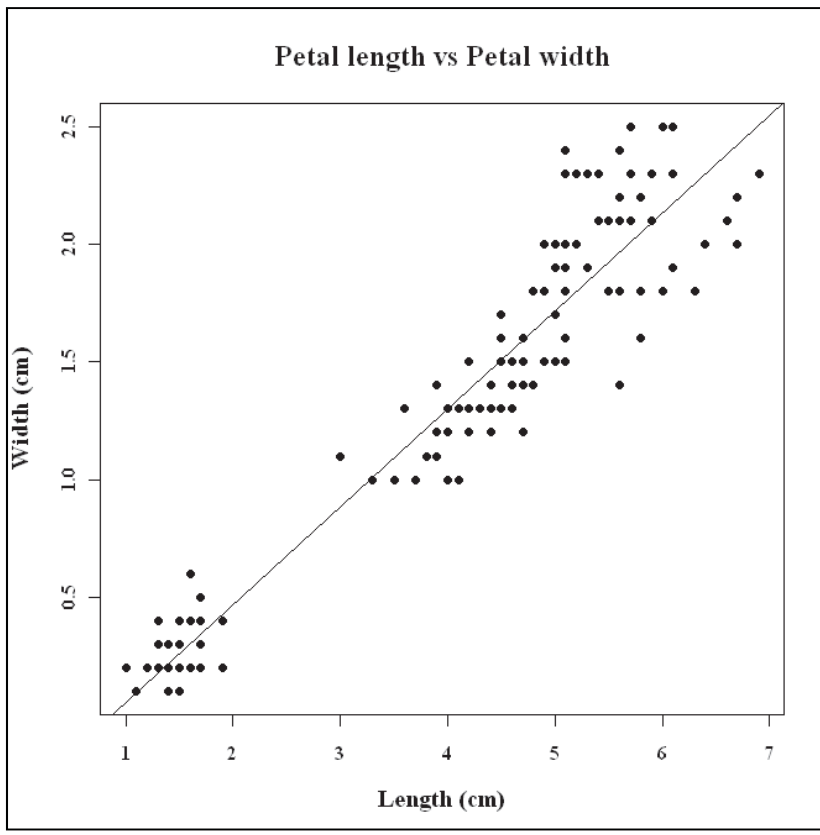
```
>plot(Petal.Length,Petal.Width,pch=19,xlab="Length (cm)",ylab="Width
(cm)",main="Petal length vs Petal
width",font.axis=2,font.lab=2,cex.main=1.5,cex.lab=1.25,family="serif")
```

Otra cosa que podría resultar interesante añadir a este gráfico sería una recta de regresión. Para ello utilizaremos la función «abline()». En esta función es necesario especificar los valores de los parámetros de la recta (en nuestro caso, al ser regresión lineal, la ordenada en el origen, argumento «a» y la pendiente, argumento «b»). En primer lugar, para conocer los valores de dichos parámetros, haremos un modelo lineal:

```
>reg<-lm(iris$Petal.Width~iris$Petal.Length)
>reg
>abline(a=-0.3631,b=0.4158)
```

```
> reg<-lm(iris$Petal.Width~iris$Petal.Length)
> reg

Call:
lm(formula = iris$Petal.Width ~ iris$Petal.Length)

Coefficients:
      (Intercept)  iris$Petal.Length
          -0.3631             0.4158

> abline(a=-0.3631,b=0.4158)
```

Así daríamos por terminada esta gráfica, cuyo aspecto final sería el siguiente:



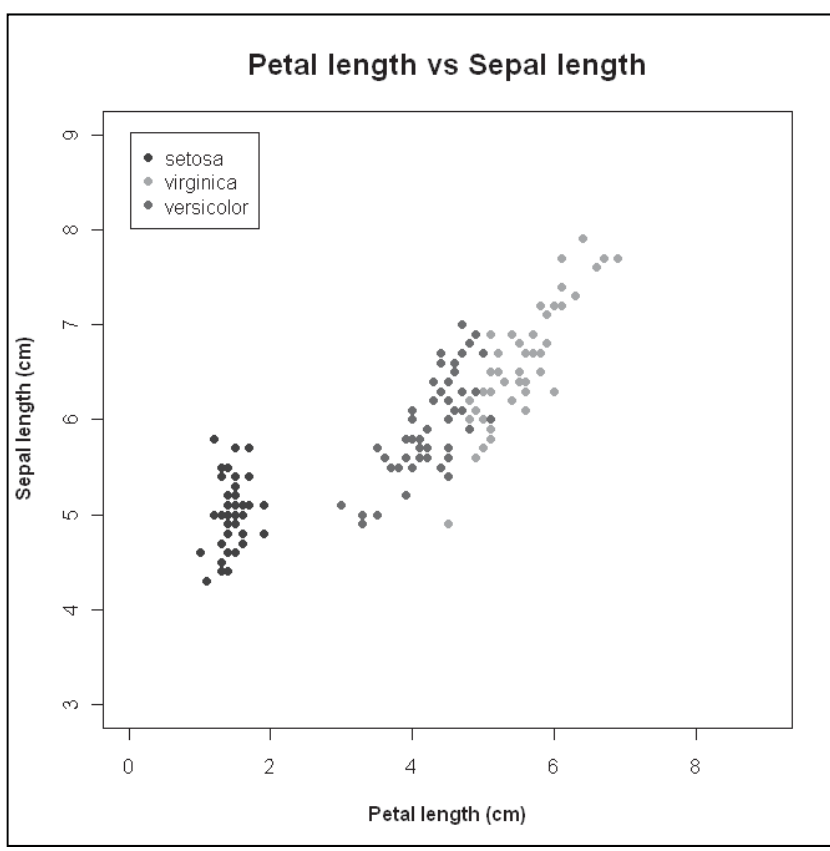
Vamos ahora con un ejemplo en el que complicaremos un poco más las cosas. Se trata de una gráfica de dispersión similar a la anterior, en la que queremos resaltar diferentes grupos de datos. Las instrucciones están en el archivo **AII.2.R**.

Utilizaremos nuevamente el conjunto de datos iris, por lo que debemos cargarlo con «data» y adjuntarlo con «attach» para poder localizar a las variables por su nombre. Vamos a representar la variable longitud de pétalo frente a la variable longitud de sépalo, diferenciando los resultados para cada especie. El conjunto total de órdenes se muestran a continuación; iremos explicándolas una por una.

La primera instrucción, «plot», sirve para representar el diagrama de dispersión para la especie *setosa*. Además añadimos un rango de valores para los ejes, «xlim» e «ylim», indicamos que se usa un círculo relleno como símbolo «pch», de color azul «col» y que la etiqueta de los ejes esté en blanco (eso es así porque está previsto etiquetar los ejes con otra función). En las siguientes dos órdenes «points» añadimos los valores correspondientes a las otras dos especies, con el mismo símbolo en un color diferente, verde y rojo (para los colores podríamos usar un código en números además del código en letra). En la siguiente instrucción utilizamos la función «title()» para añadir el título del gráfico y las etiquetas de los ejes, especificando que las fuentes de ambos deben estar en cursiva y el tamaño del título aumentar en un 50%. Por último añadimos un recuadro con la

leyenda (función «legend()») que situaremos sobre el gráfico con la función «locator()». Esta función permite obtener las coordenadas de cualquier punto que se señale sobre el plano de la gráfica con el ratón y situar el ángulo superior izquierdo de la leyenda sobre él. En lugar de la función «locator()» también podríamos escribir las coordenadas directamente. El siguiente argumento indica los nombres que debe llevar la leyenda y después los símbolos y colores respectivos. Al ejecutar esta última orden, la leyenda no aparece hasta que hagamos click en el lugar en que deseemos situarla (debido a la función «locator»).

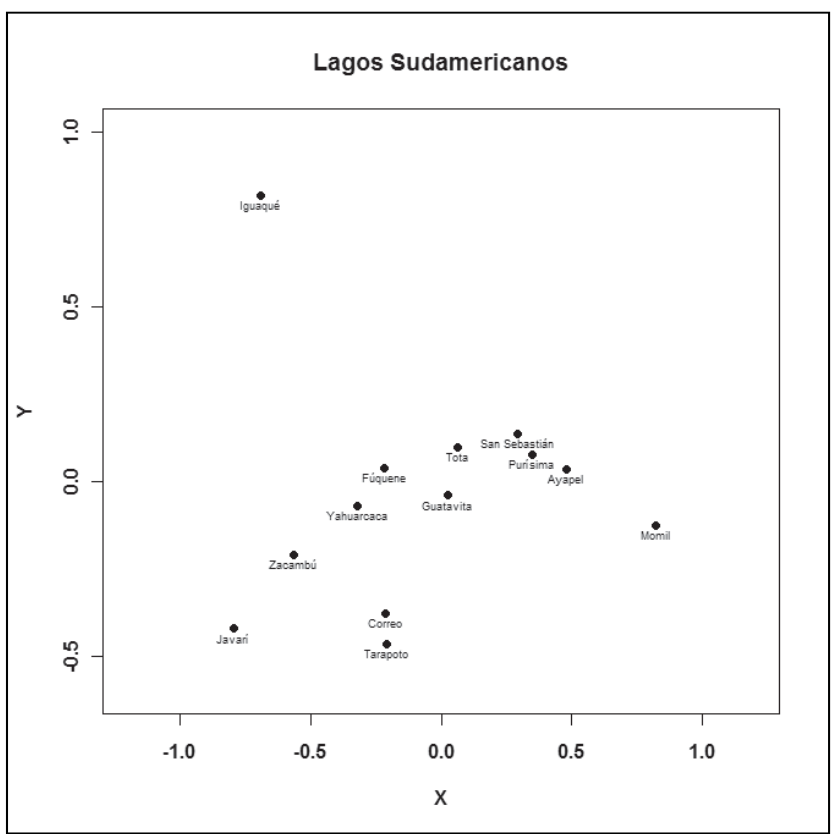
```
>data(iris)
>attach(iris)
>plot(iris[Species=="setosa","Petal.Length"],iris[Species=="setosa","Sepal.Len
gth"],xlim=c(0,9),ylim=c(3,9),pch=19,col="blue",xlab="",ylab="")
>points(iris[Species=="virginica","Petal.Length"],iris[Species=="virginica","Se
pal.Length"],pch=19,col="green")
>points(iris[Species=="versicolor","Petal.Length"],iris[Species=="versicolor","
Sepal.Length"],pch=19,col="red")
>title(main="Petal length vs Sepal length",xlab="Petal length (cm)",ylab="Sepal
length (cm)",font.main=2,font.lab=2,cex.main=1.5)
>legend(locator(1),c("setosa","virginica","versicolor"),pch=c(19,19,19),col=c("b
lue","green","red"))
```



Hasta ahora hemos utilizado la función «plot» para representar puntos en un gráfico de dispersión, pero también es posible usarla con otros símbolos, como puntos unidos por líneas o líneas simplemente. Solo habrá que indicarlo así en los argumentos de la función, como veremos en el próximo ejemplo.

En el siguiente ejemplo de gráfico de dispersión identificaremos cada uno de los puntos. Las instrucciones están en el archivo **AII.3.R**. Los datos consisten en las coordenadas obtenidas a partir de un método de ordenación espacial realizado con ciertas características de las comunidades fitoplanctónicas de una serie de lagos sudamericanos. Para ello representamos una gráfica de dispersión simple entre las coordenadas *x* e *y*. Posteriormente, por medio de la función «text()» etiquetamos cada una de las coordenadas del gráfico anterior con los valores de la variable «Lago» que se encuentran en el archivo original, es decir, el nombre de cada lago. El argumento «pos» indica la posición en un eje lateral de la etiqueta respecto al punto (el valor 1 significa centrado), el argumento «offset» indica la posición en el eje vertical respecto al punto, y el argumento «cex» indica el tamaño de las etiquetas.

```
> datos<-read.csv2("Ordenación Lagos.csv",header=TRUE, encoding= "latin1")
> attach(datos)
> plot(X,Y,xlim=c(-1.2,1.2), ylim=c(-0.6,1), main="Lagos Sudamericanos",
font.main=2,font.lab=2,font.axis=2,pch=19)
> text(X,Y,labels=Lago,pos=1,offset=0.2,cex=0.6)
```

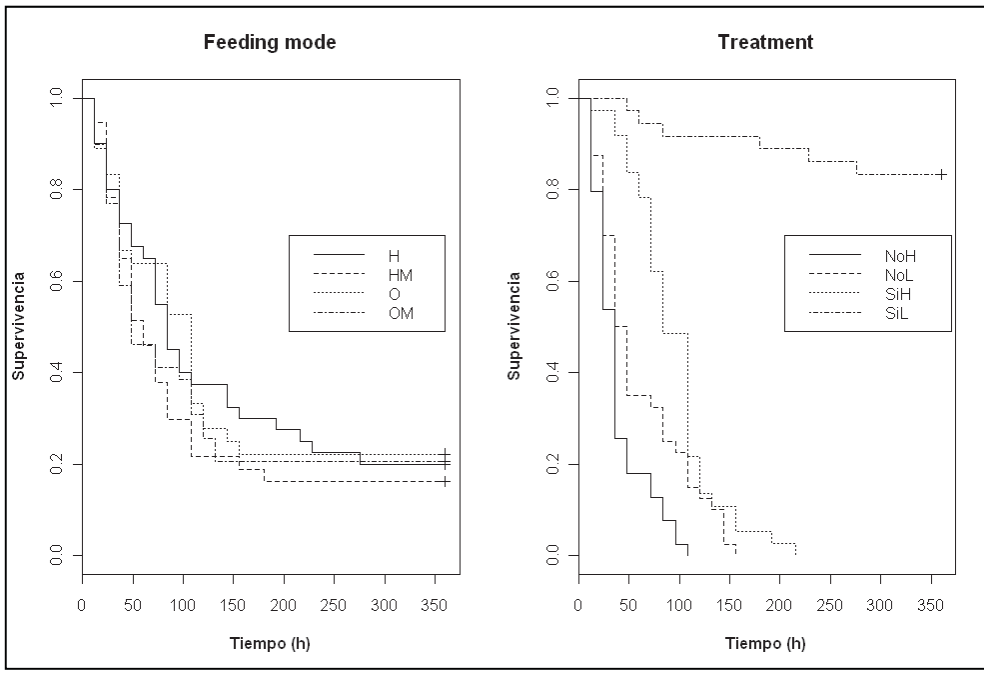


AII.4.14.3. Representar modelos y gráficos en paneles

Ahora veremos un ejemplo en el que representaremos un modelo de supervivencia. Las instrucciones están en el archivo **AII.4.R**. Para esto también nos sirve la función «plot()», gracias a que con ella es posible representar ciertos objetos como modelos. Los datos consisten en la supervivencia de individuos de cladócero de la especie *Daphnia sp*, los cuales han sido sometidos a diferentes tratamientos en los que cambia el tipo de alimento, la concentración, y si reciben o no un tipo de suplemento alimenticio. Haremos dos modelos; en el primero de ellos (modelo1) ajustamos la supervivencia al tipo de alimento que reciben «*FeedMode*» y en el otro (modelo2) tendremos en cuenta si reciben o no el suplemento y las diferencias de concentración. Queremos que los modelos sean representados por líneas y en el mismo panel queremos un gráfico para el modelo1 y otro para el modelo 2.

En primer lugar se carga el paquete "survival", que contiene las funciones necesarias para trabajar con modelos de supervivencia. A continuación se carga el archivo de datos y se adjuntan los datos. Después ajustamos los modelos con la función «survfit()», que usa «Surv()» para definir el modelo. Después utilizamos la función «par()» para definir diferentes características de los gráficos. En este caso, «mfrow()» sirve para indicar que queremos un panel con varias gráficas, cuyos argumentos son el número de filas y de columnas que queremos. En nuestro caso queremos una fila y dos columnas, así que en el panel se dibujarán dos gráficas. A partir de ahí, ya sólo quedaría crear la gráfica con la función «plot()». En este caso, queremos que se represente con líneas, por lo que utilizaremos el argumento «lty», en el que se indica qué tipo de línea representará cada grupo de datos. Los tipos de línea están codificados con números. En nuestro caso indicamos que se representen las líneas tipo 1 a 4. Después añadimos títulos y rótulos, configuramos el formato de algunas características de la fuente, y por último una leyenda, similar a la del ejemplo anterior. Para añadir la leyenda, debemos saber el orden de los tratamientos, para indicar a cuál le corresponde cada tipo de línea. En este caso, para situar ambas leyendas exactamente en el mismo punto del área del gráfico, utilizamos coordenadas en lugar de la función «locator» del caso anterior.

```
> library(survival)
> datos<-read.csv2("DaphniaSurv.csv",header=TRUE,encoding="latin1")
> attach(datos)
> modelo1<-survfit(Surv(Time,Status)~FeedMode)
> modelo2<-survfit(Surv(Time,Status)~Trat+Conc)
> par(mfrow=c(1,2))
> plot(modelo1,lty=1:4,main="Feeding mode",xlab="Tiempo
(h)",ylab="Supervivencia",font.main=2,font.lab=2,cex.main=1.25)
> legend(205,0.7,c("H","HM","O","OM"),lty=1:4)
> plot(modelo2,lty=1:4,main="Treatment",xlab="Tiempo
(h)",ylab="Supervivencia",font.main=2,font.lab=2,cex.main=1.25)
> legend(205,0.7,c("NoH","NoL","SiH","SiL"),lty=1:4)
```

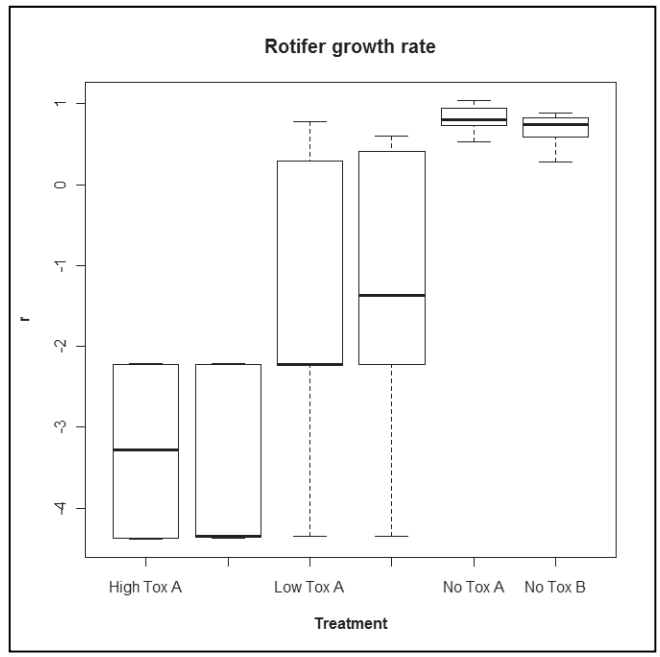


AII.4.14.4. Diagrama de cajas

Un tipo de gráficos muy utilizado es el diagrama de cajas. En él se representan los datos en función de su mediana (línea gruesa dentro de la caja) y otros parámetros relacionados con su dispersión.

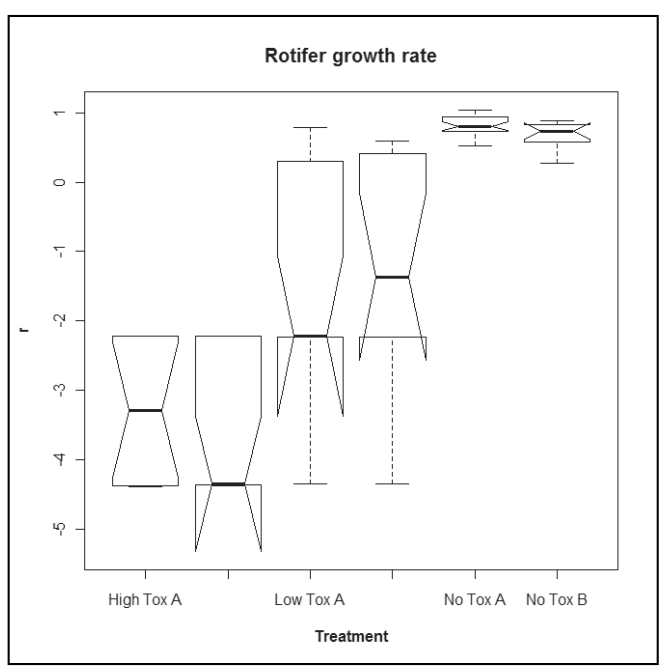
La caja abarca el rango en el que se encuentran el 1er y 3er cuartil. Las líneas punteadas indican el máximo y el mínimo si no hay outliers, o 1,5 veces la longitud de la caja si los hay. Los *outliers* (valores atípicos o extremos) se representan como puntos aislados más allá de los límites de la línea punteada. Aquí representaríamos este tipo de gráfico, con datos de la tasa de crecimiento de rotíferos a los que se somete a diferentes tratamientos, algunos de ellos con alimento tóxico. Las instrucciones están en el archivo **AII.5.R**.

```
>datos<-read.csv2("Rotíferos.csv",header=TRUE,encoding="latin1")
>attach(datos)
>boxplot(r~Treatment)
>title("Rotifer growth rate",xlab="Treatment", ylab="r",font.main=2,
font.lab=2,cex.main=1.25)
```



Una versión de este tipo de gráfico consiste en añadir un elemento, llamado «notch» (muesca) que representa intervalos de confianza aproximados al 95% para la mediana: dos medianas son distintas (los grupos son distintos) si los intervalos o muescas de las cajas correspondientes no se solapan.

```
>boxplot(r~Treatment,notch=TRUE)
```

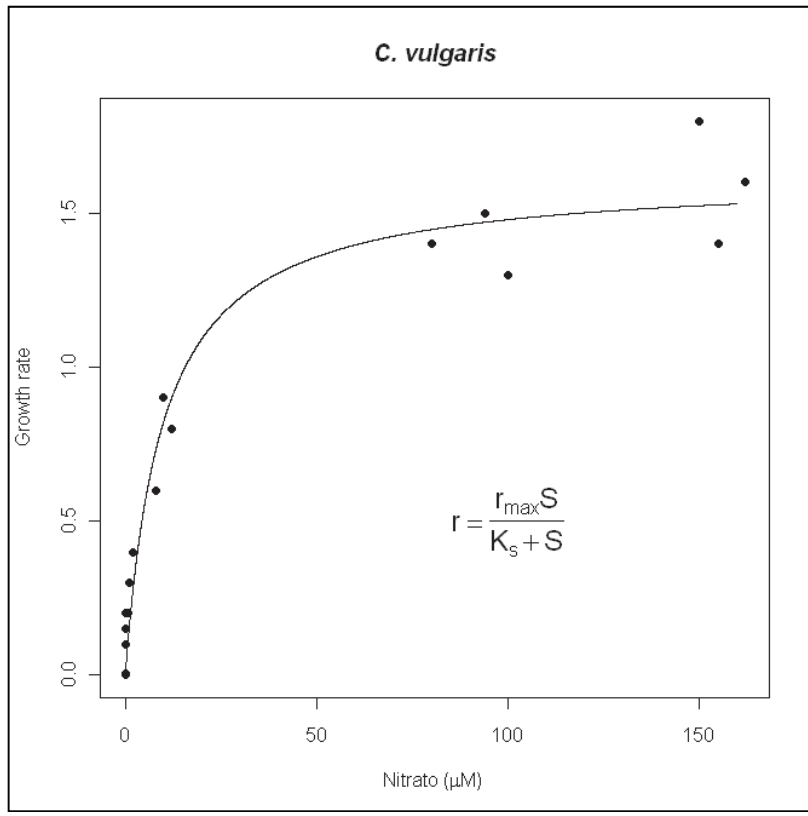


AII.4.14.5. Graficar curvas de ajuste y caracteres especiales

Otro tipo de gráficos de uso muy frecuente son aquellos en los que se representa una curva ajustada a un diagrama de dispersión de datos. La curva puede ajustarse por métodos paramétricos o no paramétricos. En nuestro caso veremos un ejemplo paramétrico. Se trata de los resultados de tasa de crecimiento de una microalga (*Chlorella vulgaris*) con nitrato como recurso limitante. Queremos representar una curva del modelo de Michaelis-Menten. Para ello debemos crear por separado un objeto que sería la curva predicha por el modelo sobre unas coordenadas determinadas, que añadiremos a nuestro gráfico con la función «lines()». Aprovecharemos este ejemplo para mostrar cómo se pueden escribir caracteres especiales en la leyenda de los ejes y fórmulas matemáticas (o cualquier otro tipo de texto) en el panel. Las instrucciones están en el archivo **AII.6.R**.

Después de cargar el archivo de datos y adjuntarlos, utilizamos la función «plot()», a la cual le añadimos el título del gráfico (esta vez en negrita y cursiva, código 4) y las etiquetas de los ejes. En este caso, la complicación está en el eje x, que debe utilizar la letra griega *μ* para indicar las unidades micromolares. Esta letra existe con el nombre de *mu* en la fuente *symbol*. Generalmente, todas las letras griegas se pueden utilizar de la misma manera (*pi*, *sigma*, *phi*, etc.). Después ajustamos un modelo de Michaelis-Menten mediante la función «nls()» y creamos dos vectores de coordenadas para representar esa función, «coordx» y «coordy». Ambos vectores deben tener la misma longitud, y en nuestro caso nos interesa que se extiendan en el mismo rango que nuestros datos. La función «lines()» dibujará las líneas de la ecuación predicha para el modelo por la función «predict()». Estas líneas no son continuas, sino segmentos entre cada par de coordenadas. Por eso al crear los vectores de coordenadas hemos definido un paso muy pequeño (0,01); así los segmentos no se notan y parece una línea continua. Por último, escribimos la ecuación de Michaelis–Menten del modelo sobre unas coordenadas determinadas (en este caso 100 y 0,5). Para ello, se utiliza la función para texto «paste()» combinada con las funciones «substitute()» y «frac()» para expresiones matemáticas, además de una serie de símbolos («[]» indica subíndice). En la página de ayuda que aparece al teclear help(symbol)» o «help(plotmath)», podemos ver una relación completa de los símbolos matemáticos disponibles para esta fuente. Además utilizamos el argumento «cex» para aumentar el tamaño de esta expresión. Es importante resaltar que los diferentes tipos de fuente (negrita, cursiva, negrita cursiva) no se aplican a las expresiones matemáticas ni símbolos de este tipo.

```
> datos<-read.csv2("Curva crecimiento.csv",header=TRUE, encoding= "latin1")
> attach(datos)
> plot(N,r,main="C. vulgaris",
xlab=expression(paste("Nitrato (",paste(mu,"M)"),sep=" ")),
ylab="Growth rate",font.main=4,cex.main=1.25,pch=19)
> modeloMM<-nls(r~rmax*N/(Ks+N),start=list(rmax=1.8,Ks=5))
> coordx<-seq(0,160,0.01)
> coordy<-predict(modeloMM,list(N=coordx))
> lines(coordx,coordy)
> text(100,0.5,substitute("r"==paste(frac(paste("r"[max]*"S"),
paste("K"[s]+"S")))),cex=1.5)
```

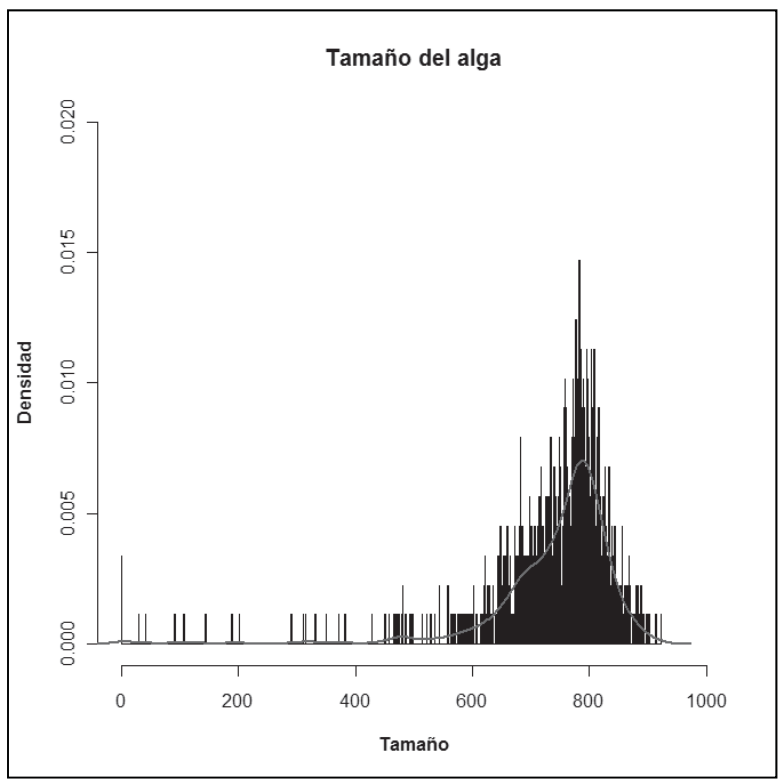


AII.4.14.6. Histograma

Ahora veremos un ejemplo de histograma sobre el que representaremos una curva de densidad de probabilidad. Jugando con los argumentos de la función para crear histogramas definiremos las clases de forma que cada barra represente –en lo posible– la frecuencia absoluta o relativa de solamente un valor. Además, introduciremos algunos nuevos parámetros gráficos. Las instrucciones están en el archivo **AII.7.R**.

Los datos de que disponemos son una serie de parámetros celulares correspondientes a una población de microalgas. Las unidades de medida se corresponden con una señal eléctrica de un determinado detector en un citómetro de flujo. Nos interesa representar un histograma sin agrupar los tamaños en clases, sino que a cada valor individual de tamaño le corresponda su densidad de probabilidad en el eje y. Con la función «hist()» representaremos el histograma. El argumento «breaks» define el número de barras del histograma y, por ello, el rango de valores de nuestra variable que abarca cada barra. Nosotros le indicamos que sea igual a la longitud de nuestra variable, por lo que cada barra abarcará un solo valor. El argumento «freq» es específico de esta función. Por defecto tiene valor «TRUE» y entonces se representarían las frecuencias absolutas de cada valor. Al darle valor «FALSE», se representan valores de densidad de probabilidad. Además, utilizamos la función «lines()» para añadir una línea que estime la densidad de probabilidad, de un modo no paramétrico.

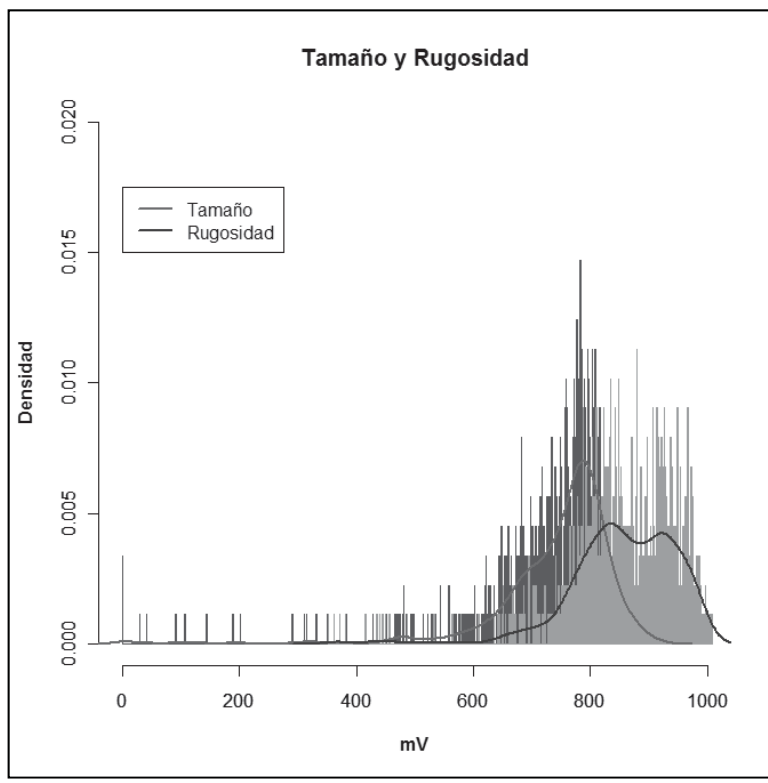
```
> datos<-read.csv2("Microalga.csv",header=TRUE,encoding="latin1")
> attach(datos)
> hist(Tamaño,breaks=length(Tamaño),xlim=c(0,1000),ylim=c(0,0.02),
main="Tamaño del alga",xlab="Tamaño",ylab="Densidad" ,font.main=2,
font.lab =2,cex.main=1.25,freq=FALSE)
> lines(density(Tamaño,na.rm=TRUE),lwd=2,col="red")
```



También es posible representar dos series de datos en la misma gráfica. En el siguiente ejemplo, cuyas instrucciones están en el archivo **AII.8.R**, además de representar el tamaño se representa la rugosidad; ambas variables se miden en mV y tienen un rango similar. Una instrucción nueva es «border», con la cual asignamos dos tonalidades de gris diferentes para el tamaño («grey22») y la rugosidad («grey55»). Utilizando la función para modificar parámetros de gráficos «par()», indicamos por medio del argumento «new», con valor «TRUE», que representaremos un nuevo histograma en el mismo panel. Al final del script también se añade una leyenda.

```
> datos<-read.csv2("Microalga.csv",header=TRUE,encoding="latin1")
> attach(datos)
> hist(Tamaño,breaks=length(Tamaño),xlim=c(0,1000),ylim=c(0,0.02),
main="Tamaño y Rugosidad",xlab="mV",ylab="Densidad",font.main=2,
font.lab=2,cex.main=1.25,freq=FALSE,border="grey22")
> par(new=TRUE)
```

```
> hist(Rugosidad,breaks=length(Tamaño),xlim=c(0,1000),ylim=c(0,0.02),
main="",xlab="",ylab="",freq=FALSE,border="grey55")
> lines(density(Tamaño,na.rm=TRUE),lwd=2,col="red")
> lines(density(Rugosidad,na.rm=TRUE),lwd=2,col="blue")
> legend(0,0.0175,c("Tamaño","Rugosidad"),lty=c(1,1),
lwd=c(2,2),col=c("red","blue"))
```

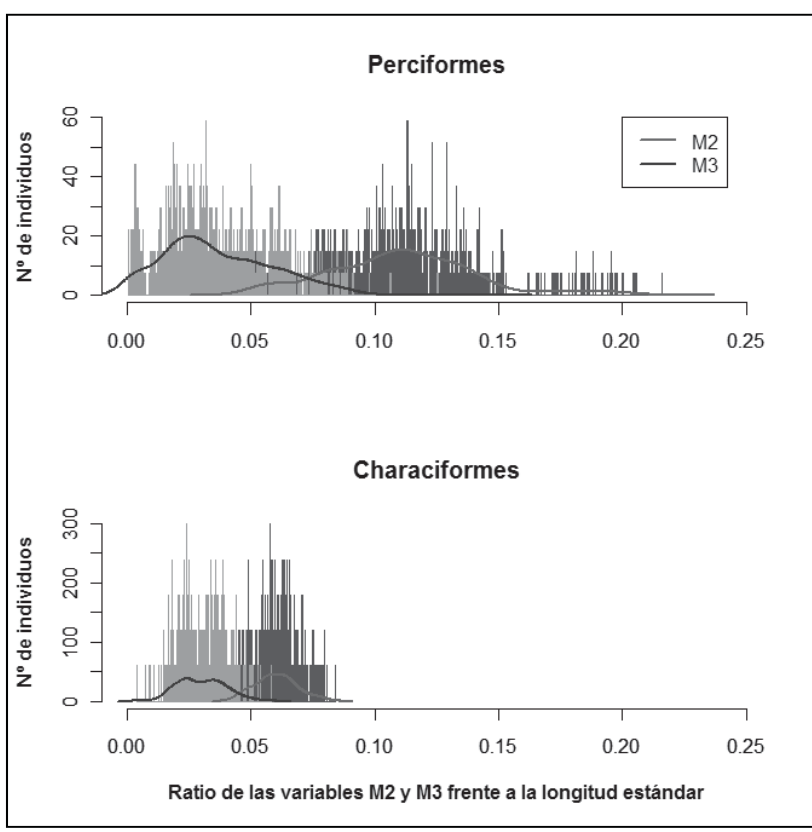


En el siguiente ejemplo se muestra cómo combinar dos histogramas en un panel y seleccionar determinados casos. Las instrucciones están en el archivo **AII.9.R**, en el cual se detalla cada uno de los pasos a modo de resumen de todo lo visto anteriormente.

Con la instrucción «par(mfrow)» se crea un panel con dos gráficas, posicionadas en dos filas y una columna. La siguiente instrucción selecciona solo el orden Perciformes. Posteriormente se hace el histograma para la variable M2. Con la instrucción «par(new=TRUE)» permitimos que el nuevo histograma con la variable M3 se haga en el mismo gráfico de antes. Luego se hacen las curvas de densidad para ambas variables y se coloca una leyenda. Posteriormente, se selecciona el orden Characiformes y se repite todo el proceso.

```
> datos<-read.csv2("Morfología.csv",header=TRUE,encoding="latin1")
> attach(datos)
> par(mfrow=c(2,1))
> datos1 <- datos[Orden=="Perciformes", ]
```

```
> hist(datos1$M2,breaks=length(M2),xlim=c(0,0.25),ylim=c(0,60),
main="Perciformes",xlab="",ylab="Nº de individuos",font.main=2,
font.lab=2,cex.main=1.25,freq=FALSE,border="grey22")
> par(new=TRUE)
> hist(datos1$M3,breaks=length(M3),xlim=c(0,0.25),ylim=c(0,60),
main="",xlab="",ylab="",freq=FALSE,border="grey55")
> lines(density(datos1$M2,na.rm=TRUE),lwd=2,col="red")
> lines(density(datos1$M3,na.rm=TRUE),lwd=2,col="blue")
> legend(0.2,60,c("M2","M3"),lty=c(1,1),
lwd=c(2,2),col=c("red","blue"))
> datos2 <- datos[Orden=="Characiformes", ]
> hist(datos2$M2,breaks=length(M2),xlim=c(0,0.25),ylim=c(0,300),main=
"Characiformes", xlab="Ratio de las variables M2 y M3 frente a la longitud
estándar",ylab="Nº de individuos",font.main=2, font.lab=2,cex.main=1.25
,freq=FALSE,border="grey22")
> par(new=TRUE)
> hist(datos2$M3,breaks=length(M3),xlim=c(0,0.25),ylim=c(0,300),
main="",xlab="",ylab="",freq=FALSE,border="grey55")
> lines(density(datos2$M2,na.rm=TRUE),lwd=2,col="red")
> lines(density(datos2$M3,na.rm=TRUE),lwd=2,col="blue")
```

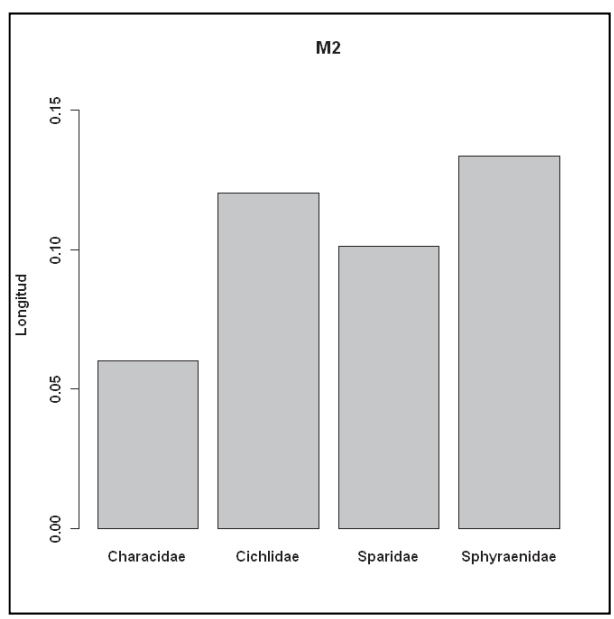


AII.4.14.7. Gráficos de barras

Además de la función «hist()», de la que ya hemos hablado, existe la función «barplot()», que nos permite representar barras simples, barras agrupadas, o barras apiladas.

Para realizar estos tipos de gráficos, partiremos del archivo que usamos en el ejemplo anterior, es decir, medidas morfológicas de una serie de peces clasificados en orden, familia, género y especie. El objetivo es representar los valores de una serie de variables por familias. Las instrucciones están en el archivo **AII.10.R**. En primer lugar, seleccionamos solamente la variable M2 para crear un gráfico de barras simples. La única función importante, que no hayamos visto hasta ahora en esta sección, es «tapply()», que usamos para obtener los valores promedio de M2 por familias.

```
#Gráfico sin representar desviaciones
> datos<-read.csv2("Morfología.csv",header=TRUE,encoding= "latin1")
> attach(datos)
> barplot(tapply(M2,Familia,mean),ylim=c(0,0.16),ylab="Longitud",
font.lab=2,font.axis=2,main="M2",font.main=2)
```

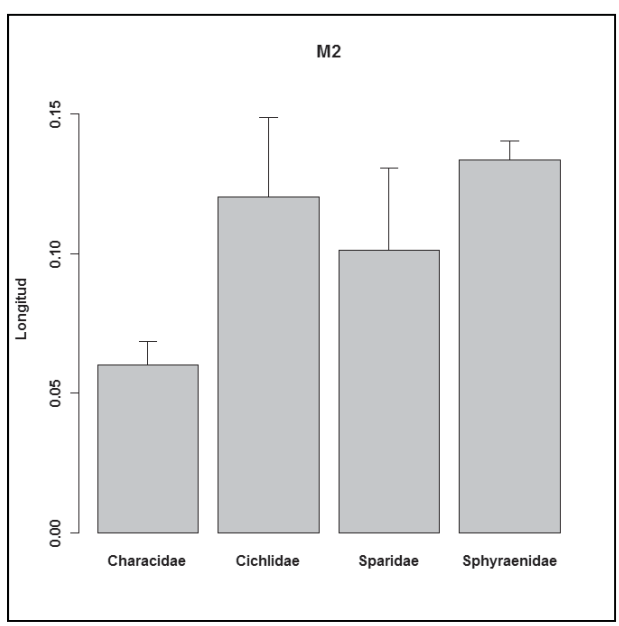


En los gráficos de barras, nos podría interesar en muchos casos representar barras de error o desviación estándar en el eje *y*. Para realizar esto no existe ninguna función *built in* que actúe directamente sobre la función «barplot()». Existen funciones como «errbar()» en el paquete "Hmisc" o «plotCI()» en el paquete "gplots" (esta última más comúnmente utilizada para representar intervalos de confianza) que pueden actuar sobre otras funciones gráficas. Por lo tanto, la mejor forma de hacerlo es crear una función definida por el usuario.

Existen muchas formas de hacerlo. Podemos, por ejemplo, utilizar un código muy simple disponible libremente en http://monkeysuncle.stanford.edu. Modifi-

caremos ligeramente este código, para que las desviaciones o errores se muestren sólo por encima de las barras y lo utilizaremos en el ejemplo de barras simples y agrupadas. En el caso de las barras simples es necesario añadir algunos detalles al código mostrado anteriormente. Destacamos que es necesario crear un objeto, «ejeX», que nos muestre los valores del eje x sobre los que se encuentran las barras (paso 1), ya que estos valores los adjudica la función «barplot» sin que nosotros los conozcamos. Creamos también en el paso 1 un objeto para los valores de eje y, «valores», además de un vector, «dvest», con las desviaciones estándar, en el paso 2.

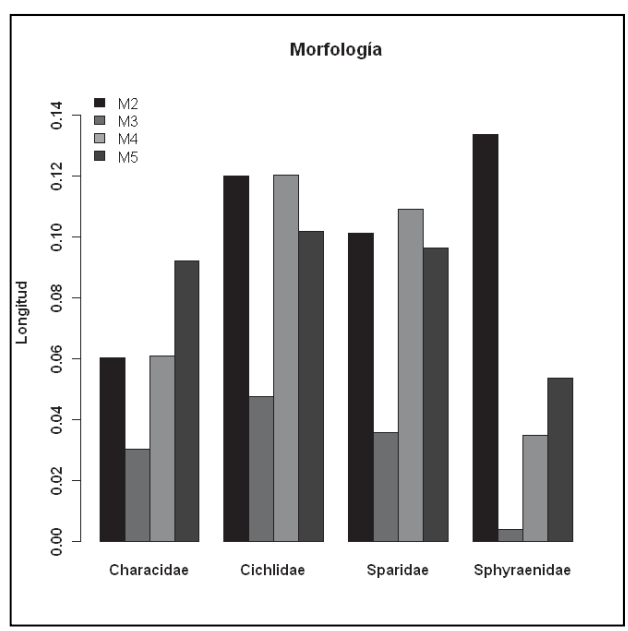
```
#Gráfico representando la desviación estándar
#Paso 1. Se representa el gráfico
> valores<-tapply(M2,Familia,mean)
> ejeX<-barplot(valores,ylim=c(0,0.16),ylab="Longitud",
font.lab=2,font.axis=2,main="M2",font.main=2)

#Paso 2. Se añade la desviación
> dvest<-tapply(M2,Familia,sd)
> error.bar <- function(x, y, upper, lower, length=0.1,...){
if(length(x) != length(y) | length(y) !=length(lower) | length(lower)
!= length(upper))
stop("vectors must be same length")
arrows(x,y+upper, x, y-lower, angle=90, code=3, length=length, ...)
}
> error.bar(ejeX,valores,upper=dvest,lower=rep(0,length=length(dvest)))
```



Ahora, representaremos los valores promedio de una serie de variables (M2, M3, M4 y M5) para cada familia, mediante un gráfico de barras agrupadas. Las

instrucciones están en el archivo **AII.11.R**. En este ejemplo ya es más complicado el manejo previo de nuestros datos, puesto que se necesitan los valores promedio de cada variable clasificados por grupos, en forma de matriz. Además, dentro de la función «barplot()», debemos indicar que el argumento «beside» tenga valor «TRUE». De no ser así, se generaría un gráfico de barras apiladas, como veremos posteriormente.

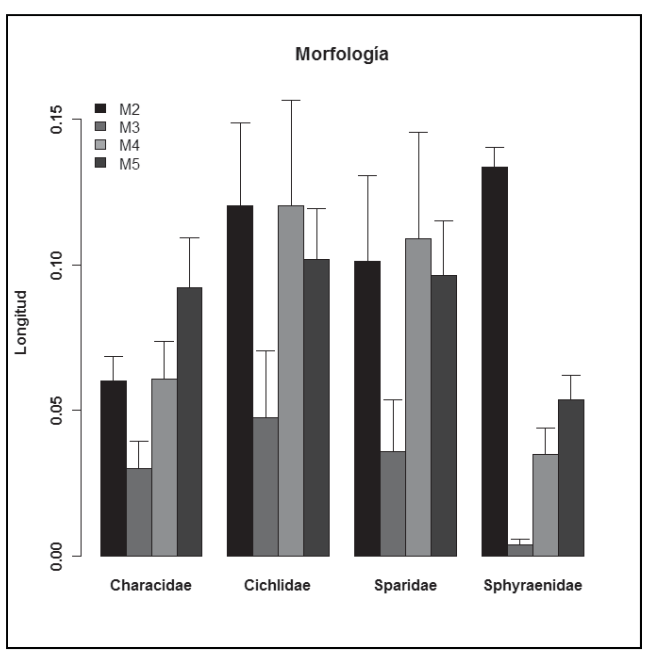
```
#Gráfico sin representar desviaciones
> datos<-read.csv2("Morfología.csv",header=TRUE,encoding= "latin1")
> datos2<-datos[,c("Familia","M2","M3","M4","M5")]
> datos2M<-aggregate(datos2[,c("M2","M3","M4","M5")],
by=list(datos2$Familia),mean)
> datos2M<-datos2M[,-1]
> rownames(datos2M)<-c("Characidae",
"Cichlidae","Sparidae","Sphyraenidae")
> datos2M<-as.matrix(datos2M)
> barplot(t(datos2M),beside=TRUE,ylim=c(0,0.15),
main="Morfología",ylab="Longitud",font.main=2,
font.axis=2,font.lab=2,col=1:4)
> legend("topleft",colnames(datos2M),bty="n",
fill=c("black","red","green","blue"))
```



Al igual que hicimos anteriormente, es posible incorporar barras de desviación en el gráfico.

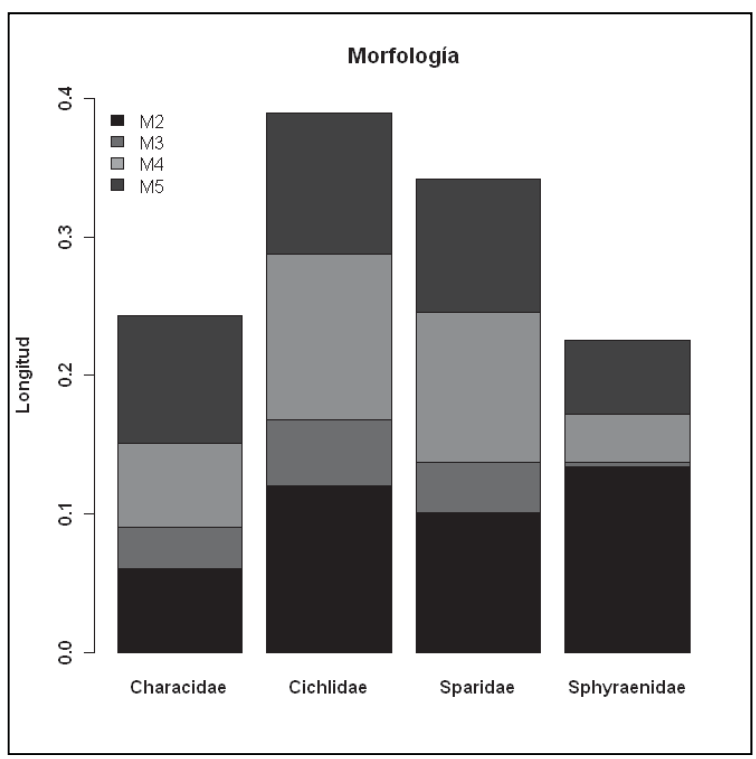
```
#Gráfico representando la desviación estándar
#Paso 1. Se representa el gráfico
> datos<-read.csv2("Morfología.csv",header=TRUE)
```

```
> datos2<-datos[,c("Familia","M2","M3","M4","M5")]
> datos2M<-aggregate(datos2[,c("M2","M3","M4","M5")],
> by=list(datos2$Familia),mean)
> datos2M<-datos2M[,-1]
> rownames(datos2M)<-c("Characidae",
"Cichlidae","Sparidae","Sphyraenidae")
> datos2M<-as.matrix(datos2M)
> ejeX<-barplot(t(datos2M),beside=TRUE,ylim=c(0,0.16),
main="Morfología",ylab="Longitud",font.main=2,
font.axis=2,font.lab=2,col=1:4)
> legend("topleft",colnames(datos2M),bty="n",
fill=c("black","red","green","blue"))

#Paso 2. Se añaden las desviaciones
> datos2Mdvest<-aggregate(datos2[,c("M2","M3","M4","M5")],
by=list(datos2$Familia),sd)
> datos2Mdvest<-datos2Mdvest[,-1]
> datos2Mdvest<-as.matrix(datos2Mdvest)
> error.bar <- function(x, y, upper, lower, length=0.1,...){
if(length(x) != length(y) | length(y) !=length(lower) | length(lower)
!= length(upper))
stop("vectors must be same length")
arrows(x,y+upper, x, y-lower, angle=90, code=3, length=length, ...)}
> error.bar(ejeX,t(datos2M),upper=t(datos2Mdvest),lower=rep(0,length=
length(datos2Mdvest)))
```



Por último, simplemente dejando el valor por defecto del argumento «beside» en el gráfico sin dispersiones, obtendríamos un gráfico de barras apiladas. Las instrucciones están en el archivo **AII.12.R**.

```
> datos<-read.csv2("Morfología.csv",header=TRUE,encoding= "latin1")
> datos2<-datos[,c("Familia","M2","M3","M4","M5")]
> datos2M<-aggregate(datos2[,c("M2","M3","M4","M5")],
by=list(datos2$Familia),mean)
> datos2M<-datos2M[,-1]
> rownames(datos2M)<-c("Characidae",
"Cichlidae","Sparidae","Sphyraenidae")
> datos2M<-as.matrix(datos2M)
> barplot(t(datos2M),ylim=c(0,0.4),
main="Morfología",ylab="Longitud",font.main=2,
font.axis=2,font.lab=2,col=1:4)
> legend("topleft",colnames(datos2M),bty="n",
fill=c("black","red","green","blue"))
```

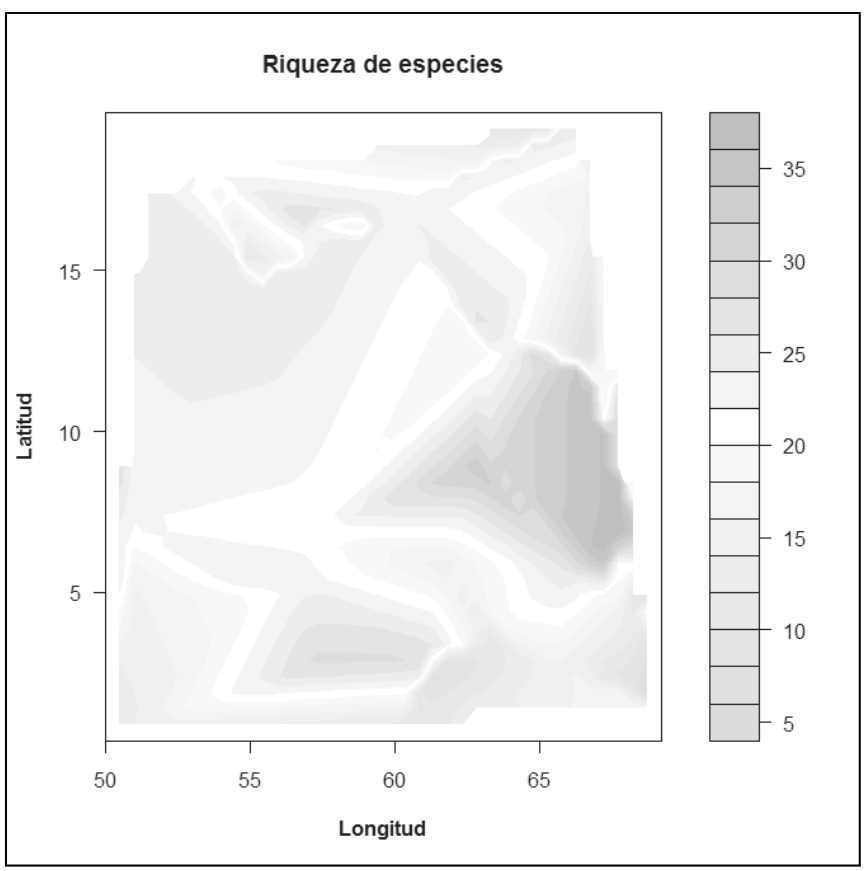


AII.4.14.8. Gráficos de contornos y superficies

Las gráficas tridimensionales son fáciles de realizar en R si disponemos de los datos adecuados. Hay varias funciones en el paquete "graphics" como «image()», «contour()» y «filled.contour()» que nos permiten realizar este tipo de gráficos. Veremos un ejemplo con esta última función. Las instrucciones están

en el archivo **AII.13.R**. En el archivo **Riqueza.csv** hay datos de número de especies para diferentes latitudes y longitudes. Para este tipo de funciones, primero tenemos que crear un objeto dividido en niveles con un intervalo regular, y con valores crecientes de las variables *x* e *y*. Si no disponemos de esos datos, como es nuestro caso, debemos generarlos. Para eso nos sirve la función «interp()», presente en el paquete "akima". Una vez que poseemos este objeto, es muy sencillo el uso de la función gráfica «filled.contour()» o similares. Sólo es importante tener en cuenta que la instrucción «interp()» no admite celdas vacías. Para poder ver los valores bajos del eje *z*, a veces es conveniente poner valores cero de *z*. Una escala de grises se puede ver con el argumento «col=grey.colors(33, start=1,end=0,gamma=1)».

```
> require(akima)
> datos<-read.csv2("Riqueza.csv",header=TRUE,encoding="latin1")
> attach(datos)
#función interp() no admite NA.
> input<-interp(Longitud.Oeste,Latitud.Sur,Riqueza.especies, duplicate=
"mean")
filled.contour(input,xlab="Longitud",ylab="Latitud",
main="Riqueza de especies",font.lab=2)
```

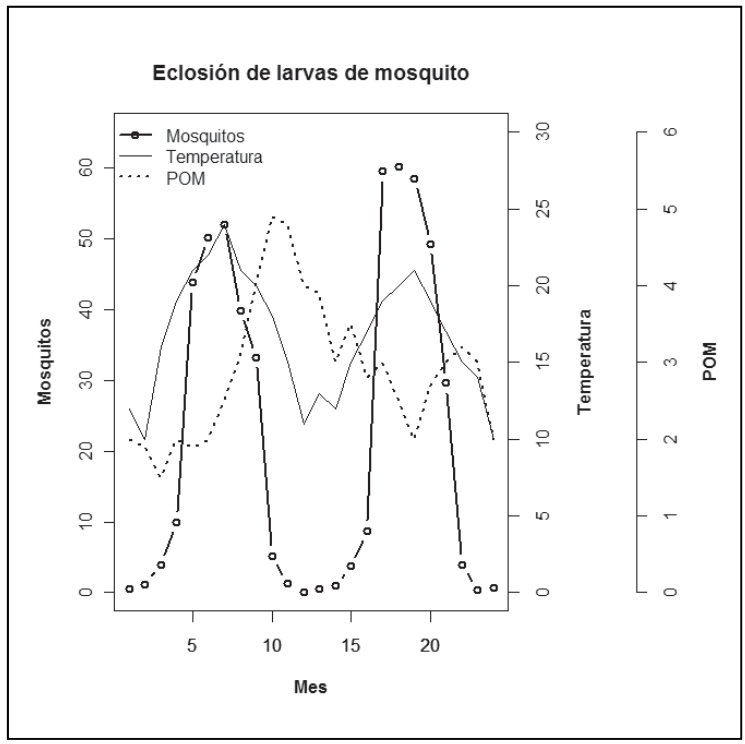


AII.4.14.9. Configuración de márgenes y ejes adicionales

Se va a utilizar un ejemplo de series temporales para mostrar como representar varios ejes *y* en el gráfico, ya que representaremos variables cuyas escalas son muy diferentes. Las instrucciones están en el archivo **AII.14.R**. Se trata de una serie temporal de medidas mensuales, de dos años de duración, realizadas en una laguna. Las variables medidas son: eclosión promedio de larvas de mosquito por metro cuadrado, materia orgánica particulada en mgl^{-1} y temperatura media del agua en el momento del muestreo. Representaremos un eje *y* distinto para cada variable.

En primer lugar debemos establecer un tamaño para los márgenes externos con la instrucción «par(oma=c())» dejando espacio suficiente para todos los ejes, especialmente del lado derecho. Después creamos los sucesivos gráficos, con la particularidad de que utilizamos el argumento «new=TRUE» antes de cada uno de los gráficos que representarán cada variable en un eje distinto. En estos nuevos gráficos, indicamos que el eje *y* no se represente «yaxt="n" ». Ese eje lo dibujaremos después con la función «axis()», indicando el margen (4 significa a la derecha). En el segundo caso (eje para POM) indicaremos que el eje lo represente en el margen exterior (argumento «outer=TRUE»), y en la línea 4. Los títulos de cada eje se añaden con la función «mtext()». En esta función, indicamos en primer lugar el texto que deseamos escribir. Además indicamos el margen en el que deseamos escribir el texto y la línea (contando desde el recuadro del gráfico) en la que se desea añadir el texto. Además, al final, añadiremos una leyenda.

```
> datos<-read.csv2("Mosquitos.csv",header=TRUE,encoding="latin1")
> attach(datos)
> par(oma=c(1,1,1,9))
> plot(Mes,Mosquitos,type="b",lty=1,lwd=2,ylim=c(0,65),
main="Eclosión de larvas de mosquito",font.main=2,font.lab=2)
> par(new=TRUE)
> plot(Mes,Temp,ylim=c(0,30),xlab="",ylab="",yaxt="n",type="l",lty=1)
> axis(4)
> mtext("Temperatura",4,line=3,font=2)
> par(new=TRUE)
> plot(Mes,POM,ylim=c(0,6),xlab="",ylab="",yaxt="n",type="l",lty=3,lwd = 2)
> axis(4,line=4,outer=TRUE)
> mtext("POM",4,line=9,font=2)
> legend("topleft",c("Mosquitos","Temperatura","POM"),bty="n",
pch=c(21,NA,NA),lty=c(1,1,3),lwd=c(2,1,2))
```



AII.4.14.10. Formas especiales

Otro aspecto interesante relativo al formato de los gráficos es la posibilidad de añadir elementos como polígonos y flechas, o texto en los márgenes. Para esto, utilizaremos como ejemplo un plano en el que se muestran las isoclinas de un modelo de competencia interespecífica de Lotka–Volterra. Las instrucciones están en el archivo **AII.15.R**.

En primer lugar crearemos un panel gráfico con los márgenes externos un poco mayores de lo habitual, para dejar sitio al texto marginal. Para esto utilizamos la función «par()», que permite ajustar parámetros gráficos, y modificamos el argumento «oma», que indica los valores de cada margen mediante un vector. Después creamos un gráfico usando la función «plot()», etiquetando los ejes en blanco. Además, no queremos que se representen los valores de las coordenadas en los ejes y, para ello, utilizamos los argumentos «xaxt» y «yaxt» con valor «n». Asimismo, tampoco queremos que se represente ningún punto en el eje, por lo que utilizamos el argumento «type» con valor «n».

```
>par(oma=c(3,3,3,3))
>plot(0,0,xlim=c(0,1),ylim=c(0,1),xlab="",
ylab="",xaxt="n",yaxt="n",type="n",
main="Isoclinas de crecimiento 0",font.main=2)
```

Ahora, mediante la función «mtext()» añadiremos texto a los márgenes. Dentro del paréntesis indicamos, en primer lugar, el texto que deseamos escribir. En este caso se trata de texto con expresiones matemáticas. Ya hemos visto

anteriormente un ejemplo similar. Además indicamos el margen en el que deseamos escribir el texto (1 indica inferior, 2 indica izquierdo), la línea (contando desde el recuadro del gráfico) en la que se desea añadir el texto, y con el argumento «at», la coordenada del gráfico a la altura de la cual queremos escribir el texto. Para localizar las coordenadas más adecuadas, resulta útil la función «locator()», que ya hemos visto, y que nos proporciona las coordenadas de cualquier punto que marquemos en el gráfico. El argumento «las», por último, indica mediante un código la orientación del texto que escribiremos.

```
>mtext(expression(K[1]),2,1,at=0.76,las=1)
>mtext(expression(frac(K[2],alpha[12])),2,1,at=0.53,las=1)
>mtext(expression(K[2]),1,1,at=0.78,las=1)
>mtext(expression(frac(K[1],alpha[21])),1,2,at=0.5,las=1)
>mtext(expression(N[1]),2,1,at=1,las=1)
>mtext(expression(N[2]),1,1,at=1,las=1)
```

A continuación nos interesa pintar en diferentes colores las distintas regiones del gráfico delimitadas por estas isoclinas. Para esto podemos utilizar la función «polygon()», que nos permite dibujar un polígono y colorearlo o rellenarlo con diferentes tramas. Dibujaremos un polígono para cada región. A esta función es necesario indicarle las coordenadas de cada vértice, lo cual se puede hacer de diferente forma según el caso. En este ejemplo se recomienda emplear la función «locator()» una vez más. Aquí, para definir bien los límites de cada polígono, cuando estos estén definidos por las líneas de las isoclinas, es bueno marcar con «locator()» encima de las isoclinas. Cuando dibujemos el polígono estas desaparecerán, ocultas por el límite del polígono. Luego se pueden volver a dibujar y los polígonos quedarán perfectamente delimitados, sin espacios en blanco. Además de las coordenadas, la función «polygon()» tiene más argumentos: en este caso definimos diferentes colores de relleno, sin trama, y no queremos dibujar bordes.

```
>polygon(x=c(-0.038,-0.038,1.039,1.039,0.78,0.255),
y=c(0.753,1.038,1.038,-0.039,-0.039,0.325),col="gray88",border=NA)
>polygon(x=c(-0.038,-0.038,0.255),
y=c(0.525,0.751,0.325),col="gray77",border=NA)
>polygon(x=c(0.255,0.78,0.494),
y=c(0.325,-0.039,-0.039),col="gray77",border=NA)
>polygon(x=c(-0.038,-0.038,0.255,0.494),
y=c(-0.039,0.525,0.325,-0.039),col="gray66",border=NA)
```

Ahora representamos unas líneas llamadas isoclinas de crecimiento 0, que definen las regiones del plano en las cuales se observa coexistencia, o dominancia de una u otra especie.

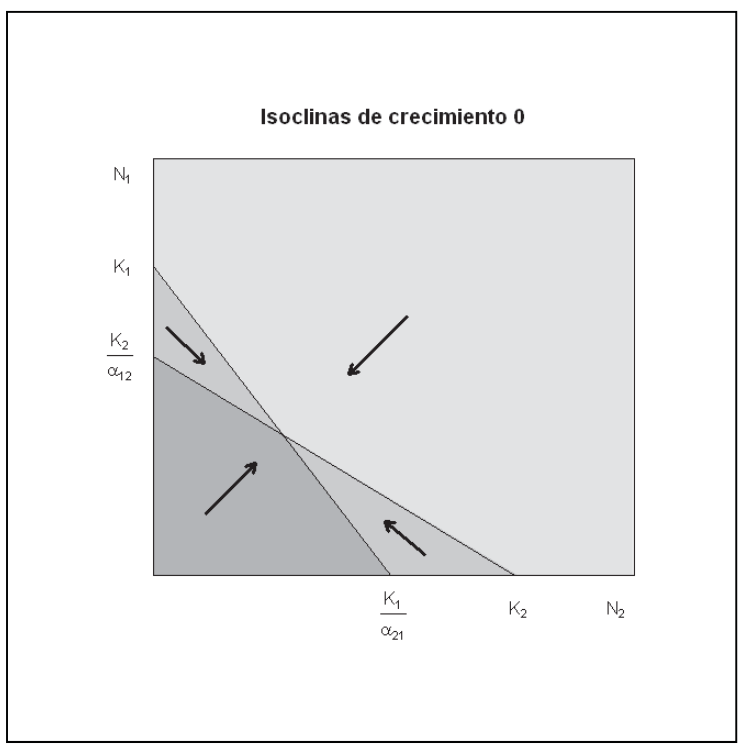
```
>abline(0.5,-0.7)
>abline(0.7,-1.5)
```

Por último, nos interesa dibujar flechas que indiquen la dirección de la resultante a la que tienden las densidades de población en cada región. En este caso todas tienden, si están más o menos centradas, al punto de equilibrio estable

definido por el cruce de ambas isoclinas. Las flechas se dibujan con la función «arrows()», a la que debemos indicar las coordenadas de inicio y de fin. Además le indicamos el tamaño de la punta de la flecha, y también un ancho de línea determinado.

```
>arrows(0.08,0.12,x1=0.19,y1=0.25,length=0.1,lwd=3)
>arrows(-0.009,0.6,x1=0.075,y1=0.51,length=0.1,lwd=3)
>arrows(0.57,0.015,x1=0.48,y1=0.1,length=0.1,lwd=3)
>arrows(0.53,0.63,x1=0.4,y1=0.48,length=0.1,lwd=3)
```

El código entero puede ejecutarse de una sola vez, obteniéndose el siguiente gráfico:



ÍNDICE DE CONCEPTOS

A

Amplitud, 30
Análisis
  de componentes principales, 667
  de correspondencias, 719
  de covarianza, 289, 318
  de varianza, 251, 318, 363
  factorial, 697
  multivariante, 667, 759
ANCOVA, 289, 318
Ángulo, coordenadas polares, 7
ANOVA, 251
  Anidado, 363
  con medidas repetidas, 318
  de Friedman, 161
  multifactorial, 252, 318
  requisitos, 217
  unifactorial, 251, 318
Antimoda, 23
Apuntamiento, 57
ARIMA, 585
Asimetría, 56
Autocorrelación, 454
Autovalor, 513, 804
Autovector, 668

B

Bondad de ajuste, 106, 107
Breslow-Day, prueba, 203
Brown-Forsythe, test, 100

C

Campana de Gauss, 49
Capacidad de carga, 520, 813
Categorías, 47, 111, 135
Centroide, 761, 776, 786, 807
Chi-cuadrado
  bondad de ajuste, 107
  contraste homogeneidad, 136
  de Pearson, 189
  distribución, 60
  independencia variables, 189
Ciclos, 637
Clasificación jerárquica, 759
Cluster, 759
Coeficiente
  τ de Kendall, 423
  de contingencia, 190
  de correlación de Pearson, 421
  de correlación parcial de Kendall, 440
  de correlación parcial de Pearson, 439
  de correlación de Spearman, 422
  de incertidumbre, 190
  de variación, 33
  Gamma (γ), 424
Colinealidad, 454
Componente
  auto regresiva, 587
  de integración, 587
  de media móvil, 587
Componentes principales, 667
Comunalidad, 698
Conglomerado
  de *k*-medias, 784
  jerárquico, 769
Contingencia, coeficiente, 190
Contraste de hipótesis, 103
  bilateral, 105
  unilateral, 105
Contribuciones
  absolutas, 720
  relativas, 720
Coordenadas polares, 1
Correlación, 421
Correlación, coeficiente, 421, 422
Correlaciones bivariadas
  simples, 421
  parciales, 439
Correspondencias, análisis, 719
Covariable, 289
Covarianza, 289, 318
Covarianza, análisis, 289, 318
Criterio
  de la varianza, 668
  de Kaiser, 668
Cuantiles, 27
Cuartiles, 27

Cuasidesviación típica, 30
Cuasivarianza, 30
Curtosis, 57
Curva de crecimiento, 520
  Gompertz, 521
  von Bertalanffy, 520
  logística, 519

D

Deciles, 27
Desviación típica, 30
Determinístico, 813
Dinámica de población, 811
Discriminante, análisis, 787
Dispersión, medidas, 29
Distribución
  asimétrica positiva, 218
  asimétrica negativa, 218
  binomial, 62
  bimodal, 23
  Chi-cuadrado, 60
  continua, 47
  discreta, 47
  F de Fisher-Snedecor, 61
  hipergeométrica, 64
  leptocúrtica, 58
  mesocúrtica, 58
  multimodal, 23
  Normal, 48
  platicúrtica, 58
  Poisson, 66
  simétrica, 56
  t de Student, 58
  unimodal, 23
Durbin-Watson, contraste, 460, 470, 480

E

Escalamiento multidimensional, 743
Ecuación
  cuadrática o parábola, 455
  curva-S, 455
  exponencial, 455
  inversa, 455
  lineal, 455
  logarítmica, 455
  logística, 519
  Monod, 521
  potencial, 455
Enlace
  completo, 760
  de centroides, 760
  simple, 760
Esfericidad, 323
Error
  estándar, 33
  tipo I, 103
  tipo II, 104
Especificidad, 697
Estacionalidad, 586
Estadística descriptiva, 21
Estandarización de datos, 2
Estocástico, 811

F

Factorial, análisis, 697
Fisher, prueba, 150
Fourier
  serie de, 648
  transformada de, 641
Friedman, ANOVA de, 161
Función
  densidad, 48
  discriminante, 792
  probabilidad, 47

G

G, prueba
  homogeneidad muestras, 136
  independencia variables, 189
Gamma (γ), coeficiente, 424
Grados de libertad, 31, 59, 136
Gráfico
  polar, 18
  puntuaciones, 669
  sedimentación, 668
Greenhouse-Geisser, corrección, 323

H

Hipótesis
  alternativa, 103
  nula, 103
Homocedasticidad de los residuos, 453
Homogeneidad de las varianzas, 94
Homogeneidad, contraste
  variables cualitativas, 135
    muestras independientes, 135
    muestras relacionadas, 161
  variables cuantitativas, 217
    muestras dependientes, 240, 318
    muestras independientes, 218, 251

Huynh-Feldt, corrección, 323, 336, 342, 350

I

Incertidumbre, coeficiente, 190
Independencia de variables
  variables cualitativas, 189
  variables cuantitativas, 421
Intervalo de confianza,
  de la diferencia de medias, 83
  de la media, 71
  de la razón de varianzas, 94
  de la varianza y desviación típica, 91
Inercia, 726, 731, 733
Intravarianza, 251

K

k-medias, clasificación, 773
Kendall, coeficiente (τ), 423
KMO, 689, 699, 705, 714, 717
Kolmogorov-Smirnov
  ajuste a una distribución de variables continuas, 108
  test de normalidad, 123
Kruskal-Wallis, prueba, 400
K-S test, 108, 123

L

Lambda de Wilks, 333, 337, 351, 354, 787
Levene, prueba, 90, 95, 99, 231, 381
Lilliefors, test, 108, 127, 268
Logística
  binomial, regresión, 537
  multinomial, regresión, 557

M

M de box, 338, 343, 357, 362, 787
Mauchley, prueba, 323, 332, 350
Mantel-Haenszel, prueba, 202, 208, 212
Matriz
  de covarianzas, 575, 668
  de correlaciones, 450, 668, 669, 672, 701, 702, 704
  de distancias,749, 751, 761, 767
McNemar, prueba, 155, 167
McQuitty, método, 761
Media
  acotada, 25
  aritmética, 21
  armónica, 22
  geométrica, 22
  ponderada, 25
  recortada, 25
Mediana, 23
Mediana, prueba de la, 405, 407
Medidas
  de dispersión, 29
  de posición central, 21
Métodos
  de clasificación, 759
  de extracción, 693, 694, 717
  de integración, 819, 829, 833, 835, 836
  de McQuitty, 761
  de Ward, 761, 771
  factoriales, 667
Moda, 23
Modelo lineal general (MGL), 251, 273, 282, 312, 347
Modelo, 811
Módulo, coordenadas polares, 1
Multivariante, 667, 759

N

Nivel significación, 103
No paramétricos, métodos, 386
Normalidad, test, 123

O

Oblimin, 698
Odds ratio, 203, 204, 208, 212, 214, 543, 555, 567
Outliers, 1, 455, 458, 459, 462, 494, 528

P

Paramétrico, métodos, 217
Percentiles, 27
Pearson, coeficiente de correlación, 421
Periodograma, 640, 649, 659
Phi, coeficiente, 203
Post hoc, pruebas
  Bonferroni, 288, 324, 333, 349, 352
  DMS, 288
  Tukey, 256, 270, 288, 289, 381, 384
Probabilidad, cálculo, 49

Promax, 698
Proximidades, 744, 754, 755, 756

Q
Q de Cochran, prueba, 179

R
Radianes, 7
Rachas, test, 387
Razón
  de varianzas, 94,
  de ventajas, 203
Recorrido intercuartílico, 33
Región
  aceptación de hipótesis nula, 104
  crítica, 104
Regresión
  variable dependiente cualitativa
    logística binomial, 537
    logística multinomial, 557
  variable dependiente cuantitativa
    múltiple lineal, 486
    no lineal, 519
    requisitos, 453
    simple, 455
Residuos, 217, 453, 454
Riesgo relativo, 203
Rotación
  equamax, 698
  oblicua, 698
  ortogonal, 698
  quartimax, 698
  varimax, 698, 703, 705, 710

S
Sattherthwaite, grados de libertad, 218
Saturaciones, 669, 675, 677, 684, 690, 694, 697, 698, 715, 718
Shapiro-Wilk, ajuste normalidad, 108
Signos, prueba, 408
Spearman (r_s), coeficiente correlación, 422
Stress, 743, 747, 752

T
t-student, distribución, 58
t-test
  muestras independientes, 218
  muestras dependientes o pareadas, 240
Tarone, prueba, 203
Test de esfericidad de Bartlett, 669, 691, 698, 717
Tukey, escalera de las transformaciones, 218

U
U de Mann-Whitney, prueba, 386

V
V de Cramer, coeficiente, 189
Validación cruzada, 789, 792, 794, 807
Variable, tipos, 47
Varianza, 30
Varianzas, homogeneidad, 94
Vectores de fuerza, 1

W
Wald-Wolfowitz, test de rachas, 387, 395, 399
Ward, método, 761, 771
Wilcoxon, prueba, 390, 399, 408, 410, 412, 414
Wilks, Lambda, 333, 337, 351, 354, 787, 805
Williams, corrección, 137

Y
Yates, corrección, 150, 210

Z
Z, estadístico distribución normal, 49